University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Barbara Agnieszka
Poszewiecka

Student no. 209493

Computational Methods for the Analysis of Chromosomal Rearrangements

PhD's dissertation

in COMPUTER SCIENCE

Supervisors:

Prof. Anna Gambin
Institute of Informatics, University of Warsaw

Krzysztof Gogolewski, PhD
Institute of Informatics, University of Warsaw

June 2023

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of PhD of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Computational Methods for the Analysis of Chromosomal Rearrangements

This dissertation focuses on algorithms for genome assembly and the interpretation of changes in genome architecture caused by structural rearrangements. The introductory chapter presents the biological background of the problem from a genetic perspective and discusses the state-of-the art sequencing technologies. The following chapter introduces an innovative algorithm designed for third-generation sequencing data. This method allows for the local assembly of regions enriched in segmental duplications. It has been utilized to reconstruct the subtelomeric sequences of selected chromosomes in Great Apes, leading to the formulation of important hypotheses about the impact of the ancestral chromosome fusion event on the evolution of pre-humans. Next, an enhanced method for determining the temporal extent of large chromosomal rearrangements is presented. This method has demonstrated its applicability in estimating the speciation times of species. The subsequent chapter introduces a method for enumerating all possible scenarios of complex chromosomal rearrangements. These rearrangements are modeled using Karyotype Graphs, which are constructed based on known breakpoints induced by the considered rearrangements. The chapter describes an algorithm for enumerating Minimal Linear Eulerian Decompositions of Karyotype Graphs, which works with a polynomial time delay complexity. Lastly, a web server enabling the clinical interpretation of structural variants is introduced. The web server features an innovative genome browser visualizing genomic regions from the rearrangement's breakpoint perspective.

Metody obliczeniowe w analizie rearanżacji chromosomowych

Niniejsza rozprawa opisuje algorytmy składania genomu oraz metody interpretacji zmian występujących w jego architekturze powodowanych przez rearanżacje strukturalne. Pracę otwiera wstęp, w którym omówiono zagadnienia biologiczne potrzebne do zrozumienia problemów przedstawionych w pracy oraz opisano nowoczesne metody sekwencjonowania genomów. Kolejny rozdział prezentuje innowacyjną metodę asemblacji genomu, która korzysta z danych z sekwencjonowania długimi odczytami. Metoda ta pozwala na lokalną asemblację rejonów wzbogaconych w segmentalne duplikacje. Wykorzystano ją do odtworzenia sekwencji subtelomerowych wybranych chromosomów u małp człekokształtnych, co pozwoliło na postawienie ważnych hipotez dotyczących wpływu zdarzenia fuzji dwóch chromosomów na ewolucję praczłowieka. Następny rozdział przedstawia opis udoskonalonej metody datowania dużych rearanżacji chromosomowych zachodzących w podczas ewolucji gatunków. Wykazano także, że można ją wykorzystać do datowania specjacji gatunków. Kolejny rozdział prezentuje metodę wyliczenia możliwych scenariuszy złożonych rearanżacji chromosomowych. Rearanżacje te modelowane są przez tak zwany graf kariotypowy, który tworzony jest na podstawie znanych punktów złamań rearanżacji. W rozdziale opisany został algorytm wyliczenia minimalnych liniowych dekompozycji Eulera w tym grafie działający z wielomianowym opóźnieniem czasowym. Ostatecznie zaprezentowano serwer umożliwiający interpretację kliniczną zmian spowodowanych przez warianty strukturalne. Udostępnia on innowacyjną przeglądarkę całogenomową pozwalającą na wizualizację obszarów sąsiadujących z punktami złamań rearanżacji chromosomowych.

Keywords

computational genomics, chromosomal rearrangements, Topologically Associating Domain, chromosomal rearrangement, position effect, clinical genetics, *de novo* assembly, long reads, complex chromosomal rearangement, enumeration, Karyotype Graph

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

J.3. Life and Medical Sciences

Tytuł pracy w języku polskim

Metody obliczeniowe w analizie rearanżacji chromosomowych

Contents

List of Figures

List of Tables

1
Introduction

Our understanding of cell biology has been revolutionized by new waves of biotechnological methods in the last decades. In particular, fascinating opportunities in the life sciences are provided by high-throughput sequencing technologies, also known as next-generation sequencing (NGS). These methods enable the parallel sequencing of multiple DNA or RNA molecules in a genome-wide, fast, cost-effective, and reproducible manner. The continuous advancement in NGS technologies allows for new applications in the field of molecular biology, including genomics (elucidating the structure, function and evolution of the genome), transcriptomics (the analysis of RNA transcripts) and epigenomics (studying phenotypic changes that do not involve alternations in the genome). To exploit their full potential in all these disciplines, novel computational methods capable of processing massive volumes of data and enabling drawing valuable biological insights are needed.

Since the beginning of the 21st century, high-throughput sequencing technologies, also known as short-read or second-generation sequencing technologies, have been commercially available. In recent years, the third-generation single molecule sequencing methods, or long-read sequencing technologies, have emerged. These technologies include Nanopore sequencing (ONT) introduced by Oxford Nanopore Technologies and single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio). The advancements in next-generation sequencing (NGS) technologies have brought significant changes to the characteristics of generated data in terms of length and accuracy, while also reducing costs. As a result, considerable progress has been made in uncovering the DNA sequence of a vast number of species, producing assemblies with dramatically improved contiguity (Hotaling *et al.*, 2021). Whole-genome sequencing (WGS), combined with novel methods for subsequent assembly and annotation, provides a more comprehensive outlook and understanding of the complete genome of an organism.

Recently, owing to rapid progress in long-read sequencing technologies, the Telomere-to-Telomere (T2T) consortium has achieved a significant milestone by releasing the first complete sequence of the human genome (Nurk *et al.*, 2022). This achievement was accomplished in noticeably shorter time and with substantially less resources compared to the initial draft human genome produced by the Human Genome Project. They have managed to close hundreds of the unresolved gaps covering highly complex regions present in the latest human reference genome (*hg38*). This feat could not have been achieved without the use of sophisticated algorithms tailored to the features of the sequencing technologies and characteristics of the complex regions of the human genome.

Noticeably, NGS data have a broad range of applications in medicine. Short-reads data are routinely used to diagnose rare diseases via sequencing of coding regions by Whole Exome Sequencing (WES) to identify variants causing rare Mendelian disorders. The advent of long-read sequencing technologies has facilitated the rapid characterization of Structural Variation (SV) in clinical settings, especially in case of balanced chromosomal rearrangements, e.g. translocation or inversions, that cannot be detected using other methods.

Combining second-generation sequencing approaches with chromosome conformation capture techniques has given raise to the Hi-C protocol. This method has led to enormous progress in elucidating the chromatin conformation. Data obtained from the Hi-C protocol are used to generate genome-wide maps of chromatin interactions at a very high resolution, allowing for the modeling of the three-dimensional architecture of the genome.

Nevertheless, all these technological advancements in NGS and its derivative technologies require the development of novel and sophisticated bioinformatic methods. These methods are crucial for determining the genome sequence and 3D structure of chromosomes, particularly in cases where they are perturbed by chromosomal rearrangements. Likewise, procedures for extracting biological and medical insights from the analysis of differences in genome architecture, at both intra- and inter- species level, need to be developed.

In this dissertation, we tackle the problem of detecting and interpreting changes in the genome architecture. The consequences of these changes can be considered both species- and individual-wise. We present two algorithms that utilize third-generation sequencing data to uncover the genome structure. Based on the outcomes of the first algorithm, that has been used for the local assembly of the fragments of the chimpanzee genome, we pose the hypothesis regarding the impact of a specific genomic event on the evolution of the ancient humans. This result supports the population genetics axiom that genome architecture plays a role in shaping species evolution. The second algorithm successfully disentangles the genome structure in patients with complex chromosomal aberrations. Additionally, we describe a web server designed to assist clinicians in leveraging information about changes in the chromosome structure. The tool enables prediction and interpretation of the clinical relevance and potential risks associated with these changes. Finally, we propose a method for the estimation of the timescope of gross evolutionary events changing the genome architecture. We believe that the presented results constitute a significant contribution to the interdisciplinary field of bioinformatics, and by making progress in the development of novel computational methods and algorithms, will help in better understanding of biological processes they describe.

As an introduction, let us provide some fundamental concepts from the field of cell genetics and biology, as well as bioinformatic terminology that will be used throughout the course of the work. Additionally, we will give a brief overview of the history and characteristics of the sequencing technologies that will help to understand the nature of the data utilized by the algorithms presented in this work.

1.1. Genetic Background

The growth, development, functioning, and reproduction of all known living organisms are controlled by genetic information stored in the *genome*. This information is encoded by units called *nucleotides*, which are composed of pentose sugar (ribose or deoxyribose), phosphate group, and one nucleobase of four types: *adenine* [A], *cytosine* [C], *guanine* [G], *thymine* [T] (or alternatively *uracil* [U]). Nucleotides are connected together into a chain polymer structure by covalent bonds, joining sugar-ring molecules of two adjacent nucleotide monomers using phosphate residues. A polynucleotide chain composed of A, C, G and U constitutes a *ribonucleic acid (RNA)* molecule. A polymer built of two polynucleotide chains composed of A, C, G and T is called *deoxyribonucleic acid (DNA)*. In DNA molecules two polynucleotide chains coil around each other using hydrogen bonds between two nucleotides forming a double helix. The binding of the nitrogenous bases of the two separate polynucleotide strands follows the pairing rule (A binds with T and C with G). The sequence of nucleobases of DNA strands determines the genetic material of an organism. Each living organism has a different DNA sequence. Human genome consists of approximately 3.2 billion nucleobases in total.

Genome of eukaryotes is composed of one or more linear chromosomes. Since chromosomal DNA can be very long, it binds to specific proteins to achieve the proper level of compaction. The DNA helix is wrapped around nucleosomes, which are composed of four pairs of histones. Due to such tight packaging this complex genome structure can fit into the cell's nucleus.

Certain genome fragments contain biological information that encodes protein production. Such a portion of an information is called a gene. Gene expression is a complex process of protein production consisting of a few steps. The first step is transcription, where the content of the gene is copied from DNA to messenger RNA molecule. Next, the messenger RNA is subject to a translation process producing a polypeptide chain, which is subsequently folded into a functional three-dimensional structure. The final step is the transfer of protein into a place in a cell where it can take part into various biochemical processes.

DNA sequences are matrices for protein production and contain elements responsible for performing structural functions or regulation of gene expression. Therefore, the identification of DNA sequences of organisms is crucial in the study of cell biology. Specifically, the ability to acquire DNA sequences has become indispensable in basic biological research, as well as other applied fields such as medical diagnosis, biotechnology, biological systematics and many more.

1.2. Sequencing Technologies

In the early 1970s, biochemists, Drs. Walter Gilbert and Frederic Sanger, devised two different methods for DNA sequencing. Gilbert's method is based on the chemical process breaking down DNA in random places. Sanger's method, on the other hand, takes advantage of the DNA synthesis reaction. In this process, a new DNA chain is synthesized base by base using sequence information on the template. The use of chemically modified nucleotides, that is dideoxynucleotides, as irreversible DNA chain terminators in Sanger's method randomly stops the synthesis process, so a series of the DNA chains of different lengths are produced. These fragments are separated in denaturing gel by electrophoresis, where smaller fragments migrate faster than larger ones. The radioactive labeling enables visualization of the fragments as bands on the gel.

Over the years, the Sanger method was further developed. The integration of the automation into the process reduced human involvement and improved efficiency. The application of fluorescently labeled terminators, instead of radioactive ones, made it safer to use and more robust. Additionally, the improved separation of DNA chains with the use of capillary electrophoresis enabled high-confidence base calls. Thanks to these developments the Sanger method became the preferred choice for the Human Genome Project. Today this method is frequently used for low-throughput DNA sequencing. With the advent of next-generation sequencing (NGS), this method is called first-generation sequencing.

Next-generation sequencing. Although it is robust in sequencing individual DNA, the Sanger method suffers from one flaw – it is relatively labor intensive. The demand for achieving high throughput has led to the development of Next Generation Sequencing (NGS) methods. All these techniques allow for sequencing entire genomes at relatively low cost, enabling hundreds of millions of DNA molecules to be sequenced at a time.

There are two major paradigms in NGS methods: Second Generation Sequencing and Third Generation Sequencing. The former, also known as short-read sequencing, provides lower-cost, higher-accuracy data that are useful for population-level research and clinical variant discovery. The latter, also known as long-read sequencing, by contrast, is well suited for *de novo* genome assembly applications.

Second Generation Sequencing. Illumina is by far the most popular NGS platform, generating the largest amount of NGS data. The process of sequencing employed by this platform consists of three steps: amplification, sequencing and analyzing. Before these steps, the DNA sample is randomly sheared into short fragments (usually around 50-500 basepairs) and ligated with adapters. Such modified DNA is loaded onto a flow cell and amplified in wells. Each of the wells contains oligonucleotides where adapters can attach. Then, each template molecule is clonally copied through the process of ”bridge amplification”. The core of Illumina sequencing technology is based on the same approach as the Sanger sequencing-by-synthesis method. What sets it apart from the Sanger method is the incorporation of specifically modified nucleotides in the way that terminator moiety only temporarily prevents the new DNA strand from expanding. After the optical detection of fluorescent label specific to the nucleotide type, the terminator moiety is cleaved, and synthesis resumes for the next cycle. After each round of synthesis, a camera takes a picture of the chip, and computer software analyzes the wavelength of the fluorescent tag determining the base for every spot on it. The process continues until the full DNA molecule is sequenced.

The overall error rate of the Illumina sequencing method is below 1% , which makes it one of the most accurate NGS platforms currently available. The most common type of error is single nucleotide substitution.

Third Generation Sequencing. Genomes are highly complex and contain long repetitive elements, copy number alterations and structural variations that are relevant in evolution, adaptation and disease. Many of these complex elements are too long to be resolved using short-read sequencing technologies. To bridge this gap, Third Generation Sequencing, also known as long-read sequencing technologies, have been developed. Reads produced by these technologies are several kilobases in length and allow for the resolution of significantly larger structural features, in many cases covering the entire complex region.

Long reads are also useful in transcriptomics research, as they are capable of spanning entire messenger RNA transcripts. Long-read sequencing technologies are offered commercially by two companies, namely by Pacific BioSciences and Oxford Nanopore Technology. Both these technologies suffer from poor sequencing accuracy compared to Illumina (90% vs. 99.9%). However, recent improvements in PacBio sequencing have led to the emergence of HiFi technology producing long reads with accuracy comparable to short-read instruments.

Pacific BioSciences (PacBio). The PacBio instruments uses Single Molecule Real Time (SMRT) technology, which enables detection of nucleotide incorporation events during the elongation of the replicated strand from the non-amplified template. The template, called a SMRTbell, is created by ligating hairpin adapters of target double-stranded DNA fragments to form closed single-stranded circular DNA. SMRTbell is loaded to a chip (SMRT cell) and immobilized at the bottom of a nanoscale unit called a zero-mode waveguide (ZMW), which provides the smallest available volume for light detection. The DNA is sequenced as the polymerase adds complementary fluorescently-labeled bases to the DNA strand by collecting light pulses emitted by laser light traveling through the glass into the ZMW. Light pulses are monitored in parallel for the primary analysis involving basecalling and adding quality values. Each SMRT cell contains an array of millions of ZMWs, capable of containing an immobilized strand of library DNA.

Pacific Biosciences (PacBio) Sequel II Sequencing instrument offers two modes of sequencing: Continuous Long Read (CLR) and Circular Consensus Sequencing (CCS, also referred as HiFi).

In CLR mode results are generated from a single continuous template from start to finish and the longest possible reads ranging from 25 Kb to 175 Kb.

Since the SMRTbell is a circular structure after the polymerase replicates one strand of the target DNA, it can continue incorporating bases of the adapter and reversed strand. If the lifetime of the polymerase is long enough, both DNA strands can be sequenced multiple times. The CSS is a multialignment of a product from single ZWM which generate a sequence with very high accuracy. The insert size in CCS mode ranges from 10 to 20 Kb.

Oxford Nanopore Technologies. All Oxford Nanopore Technology (ONT) sequencing devices use flow cells containing a nanoscale protein pore serving as a biosensor. Each nanopore is embedded in an electrically resistant polymer membrane and has its own electrode connected to a channel and sensor chip measuring the electric current that flows through the nanopore. Sequence of nucleotides is identified as it passes through such pore by measuring alteration in the ion current. Each of the four passing nucleotides causes characteristic variation in the current flow that allows for distinguishing of the nucleotides. Changes in the ionic current are decoded using basecalling algorithms allowing real-time sequencing of single molecules. The smallest ONT device, MinION, is a single flow cell containing 512 channels, with four nanopores per channel form which only one can be utilized at a time, allowing concurrent sequencing up to 512 molecules. The product for large-scale projects, PromethION, has 24 or 48 parallel flow cells incorporating up to 3,000 channels per flow cell. Currently, typical yields of MinION device are 10–15 gigabytes per flow cell, whereas for a PromethION device a yield of 153 Gb from a single flow cell has been described with an average sequencing speed of 430 bases per second (Wang *et al.*, 2021).

Although the accuracy of ONT reads is relatively low (85–94%), the read lengths mostly depend on the size of the molecules in the sequencing library and several methods have been devised for extracting and purifying high-molecular-weight DNA. Due to the improvement in the technology and library preparation the average length increased from a few thousand bases at the initial release of MinION in 2014 to 23 Kb. Recently, reads exceeding 2 Mb have been reported (Wang *et al.*, 2021).

Optical mapping. Despite recent developments in long-read sequencing methods, alternative techniques are useful to complete or confirm the order of various DNA assemblies. One of the most popular is optical genome mapping, which constructs genome-wide, high-resolution restriction maps of the DNA.

The first step in producing such maps is fragmentation of the genome into hundreds of kilobases long DNA molecules. Then, each DNA molecule is elongated on a plate. Next, it is digested using restriction enzymes and fluorescently stained at cleavage sites. The order and length of the resulting fragments are measured by imaging. Finally, the raw optical maps are *in-silico* combined into consensus molecules.

Chromosome conformation capture technologies. Chromosome conformation capture technologies are molecular biology methods for uncovering the three-dimensional organization of the genome by quantifying the number of interactions between regions that may be distant in the linear representation of the DNA molecule (Han *et al.*, 2018). All of the chromosome conformation capture-based technologies execute four main steps. First, the genome is cross-linked with formaldehyde in order to preserve fragments that are in spatial proximity. Then, these cross-linked DNA molecules are digested with the restriction enzyme. The resulting fragments are ligated to form chimeric molecules, which are subsequently reverse cross-linked to yield 3D templates. After these steps, PCR or other sequencing methods are used to quantify the frequency of interactions. There exist several techniques that differ in terms of the number of loci which interaction between can be measured.

3C (one-vs-one) The chromosome conformation capture (3C) method quantifies interactions between a pair of genomics loci, as ligated fragments are selected using PCR primers (Dekker *et al.*, 2002). This technique can be used to validate promoter-enhancer interaction.

4C (one-vs-all) Chromosome conformation capture-on-chip (4C) (also referred as circular chromosome conformation capture) measures the interactions between one genomic locus of interest (bait) and all other genomic loci (Simonis *et al.*, 2006). This method involves a second round of digestion step with a shorter restriction enzyme followed by ligation in order to produce circularized chimeric DNA fragments containing the bait. Such circularized DNA is amplified using the inverse PCR reaction. Proximity of the restriction enzyme sites to bait allows for the capturing of the ligation product of the specified DNA region. The library prepared in such a way can be hybridized to microarray or sequenced, resulting in representation of the genomic neighborhood of locus of interest.

5C (many-vs-many) Chromosome conformation capture carbon copy (5C) is an enhanced 3C method allowing for quantifying contacts between many loci simultaneously (Dostie *et al.*, 2006). The modification of the 3C method that resulted in 5C relies on changes in primer preparation method. The former technique uses specific primers, while the letter - universal ones.

Hi-C (all-vs-all) Lieberman-Aiden *et al.* (2009) have developed Hi-C protocol allowing surveying all possible pairwise interactions in fully high-throughput genome-wide manner, that effectively samples millions of interactions. The difference between Hi-C and its sister methods lies in introducing biotin into restriction enzyme cleavage sites during the DNA cross-linking. After the ligation and DNA shearing, molecules containing biotin are selected with streptavidin beads. The resulting library is pair-end sequenced to retrieve DNA fragments from each end of the ligated fragment. The results are mapped onto the reference genome and transformed to a genome-wide count matrices.

1.3. Main Results

All of the results presented in this dissertation address the problem of detecting and interpreting changes in the genome architecture shaped by chromosomal rearrangements. Each of the presented chapters describe novel bioinformatic method and their application in solving some valuable biological or clinical questions. Therefore, the results are strongly interdisciplinary, as each method is illustrated with meaningful case-study using real biomedical data.

The results in the dissertation are organized as follows.

Assembly of the regions enriched in segmental duplications

Ongoing improvements in sequencing technologies, both in terms of length and accuracy, along with continuous progress in computational assembly approaches, have enhanced our understanding of the genome architecture and evolution (Huddleston *et al.*, 2014; il Sohn and Nam, 2016; Amarasinghe *et al.*, 2020). These advancements have further reinforced the fact that segmental duplications (conventionally defined as duplicated genomic regions longer than 1000 base pairs with sequence identity greater than 90%) play a crucial role in driving structural changes and gene innovations, shaping the evolution of genomes, particularly in primates. Furthermore, according to the estimation by the Telomere-to-Telomere (T2T) Consortium, segmental duplications (SDs) constitute approximately 7% of the human genome (Vollger *et al.*, 2022).

However, *de novo* assembly of SD-rich genomic regions remains a challenging computational task due to the high error rate of NGS long reads. Currently, existing assemblers leave a significant fraction of the unassembled regions mainly corresponding to SDs when applied to long-read data significantly shorter from Ultra-Long Oxford Nanopore or with much lower accuracy than PacBio CCS (Vollger *et al.*, 2019b). A possible approach to utilize such long-read data for resolving SDs is creation of a targeted tool.

To address this challenge, we have developed PhaseDancer, a novel, fast, and robust assembler that follows a locally-targeted approach to resolve SD-rich complex genomic regions using long-read NGS data. In contrast to existing assemblers employing *top-down* paradigm operating simultaneously on all existing data, PhaseDancer produces assemblies in *bottom-up* manner gradually expanding sequence using sufficiently similar reads. Algorithm takes as an input the initial anchor sequence which is extended iteratively, by repeating four major steps. First, reads are mapped on the anchor sequence using an index of all reads loaded to RAM. In the second step, reads are clustered using randomized procedure and cluster sharing most reads with the cluster selected from the previous iteration is chosen. The third step is assembling selected reads into a contig, while fourth is extending the current anchor sequence using the contig to a new anchor sequence processed in the next iteration. After several iterations, all anchor sequences are merged to produce the final assembled sequence. Efficient integration of the state-of-the-art components used in the PhaseDancer workflow has enabled for generation of contigs with the fragments repeated up to several dozen times in the genome with at least 0.1% divergence.

PhaseDancer is additionally accompanied by the viewer and SD simulator tools. PhaseDancerViewer visualizes each step of the algorithm ie. providing a genome browser showing clusters of reads mapped on the anchor sequence. Application helps in parameter tuning and using the assembler in the semi-supervised mode, inspection of the assembly correctness, and drawing biological insights by analyzing clusters structure. PhaseDancer-

Simulator generates *in-silico* SD sequences on the basis of user-defined scenarios of their evolutionary history and, set of corresponding artificial reads with various characteristics.

To validate PhaseDancer, we have tested it on a golden-standard set of human BAC clones harboring the known SDs, and *in silico* generated SDs. PhaseDancer BAC clones assemblies have been compared to the results generated by Flye and Wtdgb2 using PacBio long-read data with 45x coverage. PhaseDancer has outclassed both general-purpose assemblers, by resolving 292 out of 341 clones (85.5%), whereas Flye and Wtdgb2 resolved 91 (26.69%) and 77 clones (22.58%), respectively. We have benchmarked PhaseDancer assembler against several commonly used assemblers supporting error-prone NGS long reads (Canu (Koren *et al.*, 2017), Wtdbg2 (Ruan and Li, 2020), Flye (Kolmogorov *et al.*, 2019), Miniasm (Li, 2016), and SDA (Vollger *et al.*, 2019a)) using evolutionary scenarios generated by PhaseDancerSimulator. From 10 evolutionary scenarios, with the number of collapsed SD copies ranging from two to twelve, PhaseDancer has resolved all of the simulated SDs with no alignment of Phred Quality Score lower than 29 (accuracy over 99.8%), while other assemblers managed to resolve at most one reference SD per scenario.

Following the successful validation of PhaseDancer, we have applied our algorithm to the unresolved subtelomeric regions of the selected chromosomes in Great Apes i.e chimpanzee, bonobo, gorilla, and orangutan, syntenic to HSA2, to unravel the mechanism of reduction of the chromosome number during human speciation after divergence from chimpanzee/bonobo. The extension of the reference sequences guided a model for HSA2 formation, resulting in the 46 chromosomes of the human species versus a karyotype of 48 chromosomes in Great Apes with a putative evolutionary advantage that might have facilitated its fixation and accumulation.

PhaseDancer assembler and model of the HSA2 formation were presented at the *26th Annual International Conference on Research in Computational Molecular Biology*, and at the *American Society of Human Genetics 2022 Annual Meeting* during poster sessions.

Methods for the estimation of the time-scope of gross evolutionary events

The reduction of the chromosome number from 48 in the Great Apes to 46 in modern humans is thought to result from the end-to-end fusion of two ancestral non-human primate chromosomes forming the human chromosome 2 (HSA2). Genomic signatures of this event are the presence of inverted telomeric repeats at the HSA2 fusion site and a block of degenerate satellite sequences that mark the remnants of the ancestral centromere. It has been estimated that this fusion arose up to 4.5 million years ago (Mya).

One of the methods of the estimation of the fusion time, proposed by Dreszer *et al.* (2007), is based on quantifying the biased gene conversions (BGCs) events. This phenomenon occurs during recombination events (Strathern *et al.*, 1995) and is a consequence of favoring strong (G, C) versus weak (A, T) nucleotide pairs at the non-Watson-Crick heterozygous sites in heteroduplex DNA in repair process (Meunier and Duret, 2004). Dreszer *et al.* (2007) has observed that BGC is locally over-represented near the telomeres of autosomal chromosomes. Using the Unexpected Bias Clustered Substitutions (UBCS) statistics measuring the bias towards weak-to-strong substitutions among the clustered substitutions and comparing their reduction in the regions near the fusion site with the orthologous telomeric sites of the chimpanzee chromosomes 2a and 2b, authors have estimated the fusion time at 0.74 Mya with a 95% confidence interval 0–2.81.

Nonetheless, the procedure of the UBCS value calculation proposed by Dreszer *et al.* (2007) is strictly constrained and considers genomic regions of the size 300 bp (window) starting every 150 bp. However, this simplification might have led to inappropriate results, especially in the subtelomeric regions containing GC-rich isochores (Costantini *et al.*, 2006). To overcome this problem, we have developed an enhanced algorithm for the re-calculation of the UBCS statistic allowing computation of its exact value for every possible window. The revised algorithm iterates over substitutions quantifying their contribution to the UBCS value of some genomic region. All windows containing each substitution are compressed into the equivalent vector of bins. Then, dynamic programming techniques are employed using formulas derived from inclusion-exclusion principle and the law of total probability to determine the exact UBCS value for one substitution. By analyzing values of UBCS statistics, quantifying the enrichment of weak-to-strong substitutions around the fusion site of HSA2, we estimated the fusion formation time at around 800,000 years ago with an upper boundary of approximately 2 Mya.

Based on the statistics derived from the enhanced algorithm, we have proposed a method for estimating speciation events times based on the average UBCS proportion between two genomes. The confidence interval has been determined using the bootstrap method for selected regions on the basis of which UBCS proportion has been calculated. The sampling has been performed on subtelomeric regions of non-acrocentric chromosomes. The speciation time has been approximated by the multiplication of the UBCS proportion by the estimated time of the human-chimpanzee split (6 Mya).

Using this method, we have reconstructed the evolutionary distances among the Great Apes (*Hominoidea*). Speciation times of chimpanzee and bonobo have been estimated very close to each other, between 4.7-6.6 Mya and 5.5-7.5 Mya, respectively. For gorilla, orangutan, and gibbon, the estimates are, respectively, 6.6-9.9 Mya, 12.5-18.4 Mya, 20.7-29.6 Mya. Noteworthy, predictions are in agreement with the literature reports (Chan *et al.*, 2010; Carbone *et al.*, 2014; Chatterjee *et al.*, 2009; Gronau *et al.*, 2011; Scally *et al.*, 2012; Stone *et al.*, 2010).

In conclusion, the results from the chapter 3 shed light on the HSA2 fusion time and provide a novel computational alternative for the estimation of the speciation chronology. The content of the chapter 3 was presented at the remote *16th International Symposium on Bioinformatics Research and Applications (ISBRA)* and published in *BMC genomics* (Poszewiecka *et al.*, 2022a).

An efficient algorithm for listing the Minimal Linear Eulerian Decompositions of the Karyotype Graphs

Complex chromosomal rearrangements are structural alterations involving more than two breakpoints. These alterations can change the orientation, order, and copy number of affected genomic segments. When they amplify genetic material or affect homologous chromosomes, the sequence of the derivative chromosomes may not be unambiguously characterized solely by the breakpoints of the rearrangements. Moreover, the presence of *de novo* copies of genomic fragments devoid of small polymorphisms makes it impossible to distinguish fragments originating from different parts of derivative chromosomes. Therefore, there is a need for an efficient algorithm that lists all possible scenarios of genomic rearrangement formation based on their breakpoints. The resulting enumeration of such scenarios can be subject to downstream analyses explaining their molecular consequences.

To address this challenge, we have devised an efficient algorithm for listing all possible scenarios of complex chromosomal rearrangements. These rearrangements can be represented by a model known as the Karyotype Graph as used in the study by Aganezov *et al.* (2019). Vertices in this graph represent the start and end points of genomic segments, while edges are of two types: segmental and adjacency. The former encode segments, while the latter encode the transitions between them. The Karyotype Graph is by definition a multigraph, and we refer to the copy number of an edge as multiplicity. The collection of trails or cycles of edges with alternating types, where the number of occurrences of each edge is equal to its multiplicity, represents an Eulerian Decomposition of the Karyotype Graph. It is a well-known fact that the Minimal (cardinality-wise) Eulerian Decomposition of a Karyotype Graph consists only of linear chromosomes if it does not contain connected components without telomeres. Each Minimal Eulerian Decomposition of such a Karyotype Graph corresponds to a rearrangement scenario.

In Chapter 4, we present an algorithm for listing all Minimal Eulerian Decompositions of Linearly Decomposable Karyotype Graphs. For this purpose, we have first reformulated this problem as the equivalent problem of listing Eulerian trails satisfying some additional properties formed by edges of alternating types in an Augmented Linearly Decomposable Karyotype Graph built upon the input Linearly Decomposable Karyotype Graph. We have approached the problem recursively, by extending the prefix of a trail in each step with edges whose decremented multiplicity (by one) does not lead to the formation of non-tivial connected components with telomeres. To this end, we have introduced two data structures: *connectivity certificate* and *witness certificate*. The *connectivity certificate* is used to determine whether decrementing the multiplicity of a given edge disconnects the graph in an undesirable way. The properties of the *witness certificate* enabled for proving that decrementing the multiplicity of only edge incident to a given vertex can form two non-trivial connected components. This has reduced the number of queries to the *connectivity certificate* during enumeration generation, resulting in significant decrease in time complexity of the algorithm. The use of the *connectivity certificate*, along with certain properties of Karyotype Graphs enable to traverse the recursion tree in a way that avoids dead ends. By employing the aforementioned ideas, our algorithm enumerates all Minimal Linear Decompositions of a Karyotype Graph with a polynomial time delay complexity of $O(log(n)^2 \cdot l)$, where n is the number of vertices in the Karyotype Graph and l is the length of the decomposition.

We have also demonstrated the utility of this algorithm in inferring plausible scenarios to explain a complex congenital rearrangement in a patient harboring such alterations.

Interpretation of structural variations disarranging 3D chromatin structure

In recent years great progress has been made in identification of SV in the human genome. However, the interpretation of variants, especially located in non-coding DNA, remains challenging. One of the reasons stems in the lack of tools exclusively designed for the clinical SVs evaluation acknowledging the 3D chromatin architecture.

To bridge that gap, we have created `TADeus2` a web server dedicated for a quick investigation of chromatin conformation changes, providing a visual framework for the interpretation of SVs affecting topologically associating domains (TADs).

`TADeus2` delivers an innovative genome browser allowing for a convenient visual inspection of SVs, both in a continuous genome view as well as from a rearrangement's breakpoint perspective. Breakpoint mode presents a track of two regions fused by the genome rearrangement, together with two tracks showing wild-type regions. It should be noted that to date this is the first genome browser with such functionality.

Importantly, `TADeus2` provides quantification and ranking of SVs pathogenicity using TADA (Hertzberg *et al.*, 2022) and ClassifyCNV (Gurbich and Ilinsky, 2020) tools. Noticeably, the second tool calculates a pathogenicity score for copy-number variant (CNV) in accordance with the American College of Medical Genetics guidelines. Additionally, an original, sampling-based method for p-value computation, quantifying the pathogenicity based on the number of disrupted enhancer-promoter interactions, has been proposed.

Furthermore, a scheme for ranking genes in the vicinity of SV according to their pathogenicity has been introduced. Ranking scheme uses the following gene characteristics: ClinGen (Rehm *et al.*, 2015) haploinsufficiency/triplosensitivity score, the number of distant candidate enhancer–promoter predicted interactions disrupted by the breakpoints of SV, number of entries of Human Phenotype Ontology (Köhler *et al.*, 2016) associated with the gene, and distance from the rearrangement breakpoints. The ranking was validated using 21 well-described cases of position effects generated by SVs and CNVs. In all cases genes contributing to the disease have been predicted as the strong (18; 85,7%) or probable (3; 14,3%) candidates.

This workflow has been successfully used in the analysis of four cases of position effects in patients carrying SVs suffering from various genetic conditions. In each of these cases putative molecular causes of syndromes have been proposed. Innovative breakpoint view has been applied in the analysis of position effect in a balanced translocation (46,XX,t(6;14)(p25.1;q12)) neighboring *FOXG1* gene in a patient with epileptic seizures and severe developmental delay. Genomic tracks showing fused regions and two wild-type regions perturbed by translocation have allowed to easily pinpoint two enhancers whose displacement lead to the abnormal phenotype. Additionally, `TADeus2` ranking scheme correctly indicated the gene responsible for the disease as a strong candidate to exhibit position effect.

Here, it should be noted that the first version of the web server was presented at the *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* and published in the conference proceedings (Poszewiecka *et al.*, 2018). The second extended version of the tool was presented in *Nucleic Acids Research* web server issue (Poszewiecka *et al.*, 2022b) and is publicly available at https://tadeus2.mimuw.edu.pl. `TADeus2` and its previous version `TADeus` was used in recently published studies (Pienkowski *et al.* (2019, 2020)).

List of publications of major results from the thesis

Poszewiecka, B., Gogolewski, K., Karolak, J.A., Stankiewicz, P. and Gambin, A., 2023. From 48 to 46 chromosomes: a novel targeted assembler of segmental duplications unravels the complexity of the HSA2 fusion. *Genome Biology*, accepted for publication by the editor.

Poszewiecka, B., Pienkowski, V.M., Nowosad, K., Robin, J.D., Gogolewski, K. and Gambin, A., 2022. TADeus2: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3D chromatin structure. *Nucleic Acids Research*, 50(W1), pp. W744–W752,

Poszewiecka, B., Gogolewski, K., Stankiewicz, P. and Gambin, A., 2022. Revised time estimation of the ancestral human chromosome 2 fusion. *BMC genomics*, 23(6), pp.1-16.

Poszewiecka, B., Stankiewicz, P., Gambin, T. and Gambin, A., 2018, December. TADeus-a tool for clinical interpretation of structural variants modifying chromatin organization. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 84-87). IEEE.

List of other publications

Pienkowski, V.M., Kucharczyk, M., Młynek, M., Szczałuba, K., Rydzanicz, M., Poszewiecka, B., Skórka, A., Sykulski, M., Biernacka, A., Koppolu, A.A. and Posmyk, R., 2019. Mapping of breakpoints in balanced chromosomal translocations by shallow whole-genome sequencing points to EFNA5, BAHD1 and PPP2R5E as novel candidates for genes causing human Mendelian disorders. *Journal of Medical Genetics*, 56(2), pp.104-112.

Murcia Pienkowski, V., Kucharczyk, M., Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., Stawiński, P., Pollak, A., Kosińska, J., Wojciechowska, K. and Lejman, M., 2020. Breakpoint mapping of symptomatic balanced translocations links the EPHA6, KLF13 and UBR3 genes to novel disease phenotype. *Journal of Clinical Medicine*, 9(5), p.1245.

Acknowledgements

First and foremost, I would like to express my deep gratitude for my supervisor, Anna Gambin, for shaping my scientific interests, helping in coordination of my research and interdisciplinary cooperation. Thank you for your tremendous support, patience, guidance, and faith during all this time.

Second, I would like to show my gratitude to Krzysztof Gogolewski, my co-supervisor, for his help, knowledge, imagination, assistance, and, most of all, friendship.

I would like to thank prof. Paweł Stankiewicz for sharing his deep knowledge, showing me the frontiers of contemporary genetics and the ways they can be extended. I am especially grateful for his supervision and care during my two scientific visits in his laboratory in Baylor College of Medicine in Houston.

Moreover, I would like to thank my scientific advisor, Michał Startek, for introducing me to the scientific world.

My thanks are also due to Maciej Sykulski, who has been a good spirit of my PhD studies, for his invaluable support, help in taking my first steps in the field of bioinformatics, and sharing his brilliant ideas.

I would like to express my deep appreciation towards my collaborators: prof. Rafał Płoski and Victor Murcia Pienkowski from the Medical University of Warsaw, and Karol Nowosad from the Medical University of Lublin for their help in taming the interdisciplinary nature of scientific problems raised in this dissertation.

I would like to show my gratitude towards Norbert Dojer and Paweł Górecki for sharing their computational resources that were crucial in conducting my research.

Furthermore, I would like to thank my colleagues: Wanda Niemyska, Agnieszka Mykowiecka, Michał Ciach, Marcin Kostecki, Błażej Miasojedow, Jarosław Paszek, Piotr Radziński, Grzegorz Skoraczyński, and Przemysław Szafrański. Without them my studies would not be such interesting and fun.

Undoubtedly, the financial support obtained from the National Science Centre was indispensable in conducting my research and preparation of this dissertation. I would like to thank this institution for founding the following grants:

- The NCN PRELUDIUM grant 2019/33/N/ST6/03110 titled ”Genome assembly algorithms for genetic disorders diagnosis”,
- The NCN HARMONIA grant 2018/30/M/NZ2/00054 titled ”Computational methods for genomic structural variants interpretation”.

Last, but not least, I would like to thank my parents, Irena and Wiesław, for their enormous support, faith in me, and understanding.

2

PhaseDancer: a novel targeted assembler of segmental duplications

"You can't make a race horse of a pig."
"No," said Samuel, "but you can make a very fast pig."
— John Steinbeck, "East of Eden"

Continuous improvement of sequencing technologies along with the development of efficient computational assembly approaches have facilitated better understanding of genome evolution and architecture (Huddleston *et al.*, 2014; il Sohn and Nam, 2016; Amarasinghe *et al.*, 2020). Segmental duplications (SDs) have been shown to be one of the key factors catalysing the dynamic evolutionary rearrangements of the genomes, particularly in primates (Marques-Bonet *et al.*, 2009a; Stankiewicz *et al.*, 2004; Ohno *et al.*, 2009). Importantly, analyses of the most recent human genome reference build (except chromosome Y) (Nurk *et al.*, 2022) by the Telomere-to-Telomere (T2T) Consortium have revealed that 7% of the human genome consists of SDs (218 Mb of 3.1 Gb) (Vollger *et al.*, 2022).

Assembly of SD-rich genomic regions has been one of the most important computational challenges in building a reference haploid genome (Vollger *et al.*, 2022). Thus far, a number of general purpose assemblers have been developed, e.g. FALCON (Chin *et al.*, 2016), Miniasm (Li, 2016), Canu (Koren *et al.*, 2017), Flye (Kolmogorov *et al.*, 2019), Wtdbg2 (Ruan and Li, 2020), Shasta (Shafin *et al.*, 2020), and HiCanu (Nurk *et al.*, 2020). Additionally, SDA assembler has been specifically dedicated to resolve SDs (Vollger *et al.*, 2019a). Currently, the high error rate of next generation sequencing (NGS) long-read data leaves a significant fraction of the unassembled regions mainly corresponding to SDs and necessitating application of targeted methods. To date, only assemblies from Ultra-Long Oxford Nanopore (UL ONT) or high-quality PacBio circular consensus sequencing (CCS) reads have been validated successfully on the data sets enriched with SDs; however, these technologies are still limited by their high cost. Technologies generating reads of length shorter than UL ONT or lower accuracy than PacBio CCS (HiFi) have turned out insufficient to accomplish these tasks (Vollger *et al.*, 2019b).

Importantly, given that the most recent T2T human genome assembly contains over 1300 SDs sites (>10 kb long, total over 227 Mb) (Fig. 2.2A) and that SD-rich human chromosome 2 (HSA2) syntenic sites in Great Apes reference genomes are incomplete (Fig. 2.2B,C), a more efficient approach to resolve their structure is needed.

We developed PhaseDancer, a novel, fast, and robust assembler that follows a locally-targeted approach to resolve SD-rich complex genomic regions. The tool is designed to work with long-reads (ONT, PacBio) and tuned for error-prone data (Fig. 2.1). Based on the iterative approach with randomised clustering procedure, the workflow of PhaseDancer enables extension of an user-provided initial sequence contig even from complex genomic regions. To assess its performance, we validated PhaseDancer using bacterial artificial chromosome (BAC) clones sampled from the known SDs as well as computationally simulated sequences reflecting a complex evolutionary history of SDs. To demonstrate efficacy and biological utility of PhaseDancer, we assembled subtelomeric regions of chromosomes 2Apter, 2Bpter, 9pter, 12pter, and 22qter in bonobo, chimp, gorilla, and orangutan together with a syntenic complex SD-rich site of HSA2 fusion that reduced the number of chromosomes from 48 in Great Apes to 46 in *Homo sapiens*, Neandertals, and Denisovans (Ventura *et al.*, 2012; Stankiewicz, 2016; Turleau *et al.*, 1972; Meyer *et al.*, 2012). Based on our assembled sequences, we have proposed a novel evolutionary model for complex HSA2 formation, indicating the most plausible key mutational events.

2.1. PhaseDancer algorithm

PhaseDancer uses an iterative greedy strategy for repetitively extending the short initial anchor sequence by executing the following phases: (i) mapping, (ii) clustering, (iii) assembling, and (iv) extending (Fig. 2.1). Additionally, we have described in details the accompanying tools: PhaseDancerViewer and PhaseDancerSimulator.

To start working with PhaseDancer, a user needs to: (i) build an index based on the read

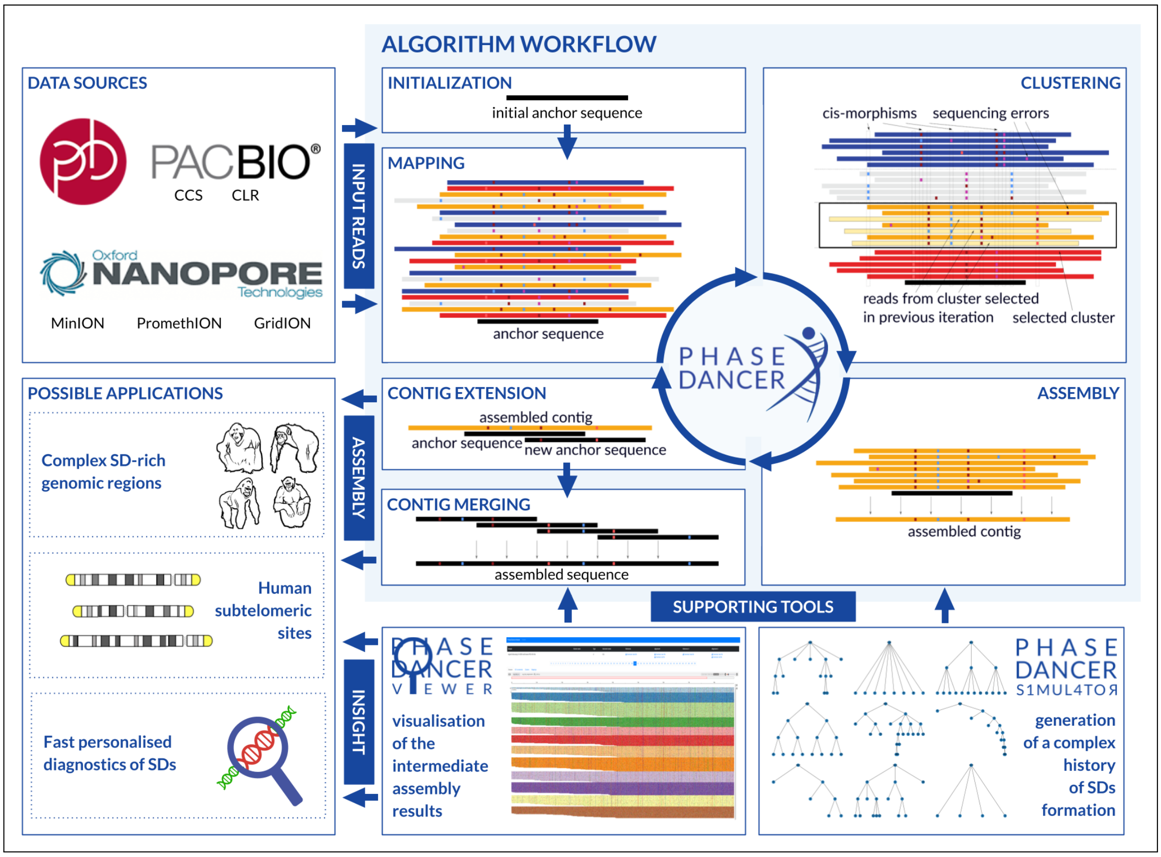


Figure 2.1: A workflow of the PhaseDancer algorithm and the accompanying tools. PhaseDancer works with next generation sequencing long-read data e.g. Oxford Nanopore or PacBio. Starting with initial anchor sequence, the core workflow of PhaseDancer iterates along four major steps: (i) mapping the reads on the anchor sequence, (ii) clustering the mapped reads and selection of a cluster with the reads originating from the genomic region represented by the anchor sequence, (iii) assembling these reads into a contig, and (iv) extending the current anchor sequence using the contig to a new anchor sequence processed in the next iteration. After all iterations, the algorithm outputs the final assembled sequence. PhaseDancer is also accompanied with two supporting tools - the semi-supervised character of PhaseDancer is provided by PhaseDancerViewer that enables the intermediate control of assembly process, whereas, PhaseDancerSimulator generates *in silico* data for profound validation of the algorithm. Thanks to its high efficiency, PhaseDancer can be used for resolving challenging genomic tasks, involving SD assembly.

data (using minimap2 (Li, 2016) tool), (ii) load the index to the RAM, and (iii) prepare the initial anchor sequences to be extended by PhaseDancer.

2.1.1. Mapping phase

An anchor sequence is mapped on the set of all reads using an inverted index loaded to RAM. Some randomly selected reads from the sample are then sent to the standard input of the mapper to load the buffer of minimap2, forcing the tool to output mappings at least once per iteration. As a result, the process of receiving the output from the mapper determines the end of the entries from the anchor sequence. The output from the anchor sequence is further processed when the first entry from the randomly selected reads is recognised.

The Pairwise mApping Format (PAF) entries generated by minimap2 are then processed to filter the reads with the sufficiently large coverage (parameterised by default with at least half of the anchor length). Selected reads are then retrieved from a FASTA file using the Faidx index.

Finally, the reads are homopolymer-compressed (HPC) and mapped on the HPC anchor sequence to produce a BAM file that is an input to the next PhaseDancer phase.

2.1.2. Clustering phase

The HPC (homopolymer-compressed) reads overlapping the full HPC anchor sequence are selected using the .bam alignment file from the previous step. Using this alignment mismatches are analysed to find candidates for cis-morphisms. Here, a cis-morphism refers to a single nucleotide difference between two or more segmental duplications.

To detect cis-morphisms, the frequency of the second most common nucleotide is computed for each locus. A locus is identified as a cis-morphism when the corresponding second most common nucleotide frequency is greater than a given threshold value (parameter dependent on sequencing technology and the coverage). Additionally, when the number of the identified cis-morphisms is greater than a given upper-bound (by default set to 200), only those with the largest percent of the second most common base are retained. Such filtered cis-morphisms are then used for clustering.

The first step of clustering is based on the graph connectivity analysis. A graph used for clustering reads is called a similarity graph. A set of vertices of the similarity graph corresponds to the reads overlapping the full HPC anchor sequence. Each edge of the similarity graph connects vertices most similar to each other according to a Hamming distance of cis-morphisms (0.4). The decomposition of the similarity graph into the connected components corresponds to the partition of the computed reads.

In the second step, each block in the partition is subdivided into clusters using cis-morphisms derived from the reads composing the block. The clustering process is based on random simulations and generates multiple alternative clusterings of reads.

In each simulation, a random cis-morphism is selected iteratively to partition the set of all reads based on the observed nucleotides. The procedure is applied recursively until either no cis-morphisms are present in the processed set of reads, or the number of reads in each constructed cluster falls below a certain threshold (this threshold is sequencing technology-dependent, yet it is assumed to be 0.8 × coverage).

Given all the alternative clusterings of reads, we assign the best clustering to each block. To evaluate the quality of a clustering, we compute the sum of distances between each read and its nearest cluster. In this context, the distance between a read and a cluster is determined by the Hamming distance between the read and the consensus sequence derived from all reads in the cluster. The best clustering is the one that minimises the sum of distances across all reads.

The final clustering of all reads is a union of all clusters from all blocks. The cluster used for the extension of the anchor sequence maximises the number of reads shared with the cluster selected for the extension in the previous iteration. In particular, in the first iteration, the cluster is selected by the similarity to the initial sequence (i.e. Hamming distance between the sequence and clusters consensuses).

2.1.3. Assembling phase

Reads from the selected cluster are pre-processed based on their mapping to the anchor sequence by truncating fragments exceeding the sequence by given flanking threshold. The procedure is applied to ensure even coverage and the fixed length of the assembly required by wtdbg2 (Ruan and Li, 2020). Then, the reads are assembled using wtdbg2. This process is fast and precise as it operates only on the reads from one cluster originating from one genomic region with the read number approximated by the coverage of the sequencing data.

2.1.4. Extending phase

The newly assembled sequence is aligned to the anchor sequence using the edlib library (Šošić and Šikić, 2017) minimising the Levenshtein distance. The flanking part is used for the extension of the current anchor sequence to the new anchor sequence processed in the next iteration (Mapping phase) of PhaseDancer.

Implementation details

PhaseDancer was implemented as a `Snakemake` (Mölder *et al.*, 2021) workflow. The source code, the docker image of PhaseDancer, and the toy-example along with the detailed manual are available at `https://github.com/bposzewiecka/phaseDancer`.

PhaseDancer uses the index of all sequencing data loaded into RAM to query for reads that are similar to the anchor sequence. Therefore, before running the main workflow the index build for all sequencing data needs to be generated. PhaseDancer uses `minimap2` `.mmi` files generated with:

- `--idx-no-seq` parameter to reduce the memory required for the index to be stored (if used, the mapper can produce an output only in the `PAF` format),
- `-p 0 -N 3000` parameters to ensure that all reads having fragments similar to the anchor sequence are outputted,
- `-K 1` parameter to force the mapper to generate an output once per read.

As a reference point for the memory usage, an index of 200 GB stored in a FASTA file uses approximately 150 GB of RAM.

Before the first iteration of PhaseDancer, the minimap2 index has to be loaded into the RAM together with two processes running in an infinite loop and handling the standard input and output of the mapper. The former receives sequences from the pipeline and sends them to the standard input of the mapper, the latter receives the output in the PAF format from the mapper, selects the reads using the Faidx index, and sends them back to the pipeline.

PhaseDancer enables the concurrent extension of many sequences. To accomplish this functionality, the input sequences are sent to the mapper using the `flock` command. Then, the process retrieving the mapping results allows for the multiplexing of PAF entries sent from many other processes. Distinction of the sender process of an entry is based on a uniquely identifying name.

2.1.5. PhaseDancerViewer - intermediate results viewer

PhaseDancer is accompanied by PhaseDancerViewer, an application for the visualisation of its intermediate assembly results obtained at the end of each algorithm iteration. The Viewer enables monitoring the assembly process in a semi-supervised mode. User can interfere the assembly process and re-tune the parameters of PhaseDancer. For every iteration, it displays the reads mapped on an anchor sequence grouped and colored by clusters. The application visualizes clusters using an embeddable implementation of the Integrative Genomics Viewer (IGV). The source code with the documentation is available at `https://github.com/bposzewiecka/phaseDancerViewer`.

2.1.6. PhaseDancerSimulator - SDs generator

PhaseDancer is targeted at resolving SD-rich genomic regions, thus the standard methods dedicated to assemblers evaluation and benchmarking are unsuitable or even inadequate. To show the advantages of PhaseDancer and verify its robustness, we implemented a simulator generating contigs and recapitulating the complex history of SDs formation.

PhaseDanceSimulator extends the method proposed by Chaisson *et al.* (Chaisson *et al.*, 2017). The simulation process follows the simplified model based on the tree topology. Fragments from a reference genome are assigned to the root of the tree and child sequences are generated by copying a parent node sequence and mutating each base at a fixed rate per base. PhaseDancerSimulator supports four topology types: flat, bifurcating, cascading, and random (Tab. 2.1). Moreover, the ends of the generated contig sequences can be extended with a randomly generated sequence.

PhaseDancerSimulator supports Oxford Nanopore and PacBio Sequel technologies using PBSIM2 (Ono *et al.*, 2021) to simulate reads. Other simulation parameters include, e.g. mutation rate, mean and standard deviation of the read length, read accuracy, chemistry, and coverage. Additionally, the tool can generate assemblies using Canu (Koren *et al.*, 2017), Wtdbg2 (Ruan and Li, 2020), Flye (Kolmogorov *et al.*, 2019), and Miniasm (Li, 2016) that can be used for benchmarking of the assemblies.

The source code and the documentation of PhaseDancerSimulator are available at `https://github.com/bposzewiecka/PhaseDancerSimulator`.

2.1.7. Runtime experiments

T2T data of SDs were used as a reference point to asses the distribution of the number of stacked SDs in the human genome needed to specify the parameters for the runtime experiments. We calculated the percent of all SD bases that have no more than n stacked SDs as: $\leq 5 \approx 65\%$; $\leq 10 \approx 79\%$; $\leq 15 \approx 83\%$; $\leq 20 \approx 87\%$; $\leq 30 \approx 90\%$. Moreover, the median number of the stacked SDs for the interstitial SDs was equal to 2. Importantly, the cases with more than 20 stacked SDs related to very short fragments.

Therefore, to conduct runtime experiments, we generated data using PhaseDancerSimulator for the number of clusters varying from 1 up to 40 (mutation rate 0.001, P6C4 PacBio chemistry, coverage 40x, sequencing error 15%, mean read length 18 kb, read length standard deviation of 3 kb, flat tree topology). The upper bound was set to 40 because in the real data scenario cases with more clusters are extremely rare, thus they do not influence the effective runtime of the algorithm.

Importantly, when assessing the runtime of the PhaseDancer number, we observed that the main bottleneck of the PhaseDancer workflow is the clustering procedure. To optimise this step, we paralleled this procedure and measured the execution time of one iteration given the number of processes used. For such generated datasets and the number of processes used (1, 5, 10, 20), we ran the experiments for 100 iterations aiming to assemble the ~ 0.5 Mb regions. To asses the time performance of PhaseDancer, one iteration time was computed for each run. The final results of the time experiments are presented in Figure 2.5A.

2.2. Methods

2.2.1. Datasets

Whole genome sequencing of two chimpanzees Using long-read PacBio Sequel II technology, we whole genome sequenced two chimpanzee (Chaos and Toby from the Houston Zoo) genomes. Chaos' genome was sequenced using CLR technology with 70x coverage, whereas Toby's genome using CCS (HiFi) technology with 20x coverage.

First, the peripheral blood DNA samples were assessed as suitable for PacBio Sequel II sequencing. DNA was fragmented with the Covaris® g-TUBE® device. Next, DNA damages were repaired using the DNA Damage Repair reagents (PacBio).

To ligate the hairpins (SMRTbell™ templates) to the DNA fragments, BLUNT hairpin adapters (20μM) oligonucleotide pre-annealed stocks) were used. To remove failed ligation products, exonuclease was added. Three-step AMPure PB Size-Selection and Purification was performed. Prior to sequencing, primer was annealed to both ends of the SMRTbell template. The binding reaction was performed and DNA sequencing polymerases were bound to the primer-annealed SMRTbell templates (at 30°C for 30 minutes).

The template-polymerase complex was transferred to a 96-well sample plate with adjusted concentrations and volumes. The DNA fragments in a zero-mode waveguide well were sequenced using PacBio Sequel II repeatedly in the sequencing process. The obtained broadcasts were self corrected to obtain highly accurate CCS reads. The resulting CCS data quality control confirmed its validity to perform the downstream analyses of the WGS from PacBio Sequel II. The P1 ratio of the two cells was over 89.62%, the average length of subreads was 14,666 bp, the read N50 was 22,239 bp, the longest read length is 268,467 bp, and the total data was 231,859,915,436 bp.

Reference genomes All reference genomes of human and Great Apes used in this study were downloaded from the UCSC Genome Browser (`https://hgdownload.soe.ucsc.edu/downloads.html`) (Kent *et al.*, 2002):

- Genome Reference Consortium Human GRCh38.p13; hg38 assembly of human genome (December 2013);
- T2T Consortium/T2T-CHM13 v2.0 assembly of the human genome (January 2022);
- University of Washington Clint_PTRv2; panTro6 assembly of the chimpanzee (*Pan troglodytes*) genome (University of Washington, January 2018));
- Chimpanzee Sequencing and Analysis Consortium Build 3.0; panTro5 assembly of the chimpanzee (*Pan troglodytes*) genome (May 2016);
- University of Washington Mhudiblu_PPA_v0 assembly; panPan3 assembly of the the bonobo (*Pan paniscus*) genome (May 2020);
- Max-Planck Institute for Evolutionary Anthropology panpan1.1; panPan2 assembly of the bonobo (*Pan paniscus*) genome (August 2015);
- University of Washington Kamilah_GGO_v0; gorGor6 assembly of the gorilla (*Gorilla Gorilla*) genome (August 2019);

- University of Washington Susie_PABv2; ponAbe3 assembly of the orangutan (*Pongo pygmaeus abelii*) genome (University of Washington, January 2018).

Great Apes NGS data from public repositories The following PacBio circular consensus sequencing (CCS) data for Great Apes were used to validate and extend the existing references:

- Chimpanzee (Clint), BioSample SAMN15896587, Bioproject PRJNA659034 (Primate genome sequencing and assembly),

- Bonobo (Mhudiblu), BioSample SAMN11123633, Bioproject PRJNA691628 (bonobo and gorilla HiFi reads),

- Gorilla (Kamilah), BioSample SAMN11078986, Bioproject PRJNA691628 (bonobo and gorilla HiFi reads),

- Orangutan (Susie), BioSample SAMN15896588, Bioproject PRJNA659034 (Primate genome sequencing and assembly).

Analysis of polymorphisms To assess the polymorphisms flanking the HSA2 fusion site, we analysed NGS data (Nanopore, PacBio CLR, and CCS HiFi) from two data sources: Genome in the Bottle (3 individuals: https://github.com/genome-in-a-bottle/giab_data_indexes) and T2T Diversity Panel (10 individuals: HG01109, HG01243, HG02080, HG03098, HG02055, HG03492, HG02723, HG02109, HG01442, HG02145, https://github.com/human-pangenomics/hpgp-data).

Optical genome mapping data All OGM data representing assembly of raw molecules in `CMAP` format were provided by Bionano Genomics and downloaded from NCBI FTP sites using URLs provided in ftp://ftp.ncbi.nlm.nih.gov/pub/supplementary_data/bionanomaps.csv. We used the optical genomic maps generated with the nicking enzymes BssSI and BspQI of the chimpanzee and orangutan genomes from the bioproject PRJNA369439, and the bonobo genome from the bioproject PRJNA672266. Gorilla Bionano Genomics data from the bioproject PRJNA369439 were generated with DLE1 nicking enzyme.

Transcriptomic analysis Bulk RNA-seq data from three species: human, bonobo, and chimpanzee available at https://www.ncbi.nlm.nih.gov under the bioproject PRJNA527986 were used to perform the comparative transcriptomic analyses of the transcriptomes from the PhaseDancer-extended subtelomeric regions.

2.2.2. Optical genome mapping validation of the recent reference genomes and PhaseDancer assemblies in Great Apes

To validate the assemblies of the reference genomes used in our work, we used the Bionano Genomics data described above. Data processing pipeline followed the producer guidelines for running Bionano Solve in the Command Line (Guidelines at https://bionanogenomics.com/).

`FASTA` files of the genome reference builds were *in silico* digested with the nicking enzymes using HybridScaffold script to produce files in the `CMAP` format. Then, the mapping was performed using the producer provided runCharacterize.py script with preset parameters optArguments_haplotype_saphyr.xml (for BssSI and BspQI enzymes) and optArguments_haplotype_DLE1_saphyr.xml (for DLE1 enzyme) accompanying the script. The produced mapping was visualised using the Bionano Access Server.

2.2.3. Bulk RNA-seq gene expression analysis

RNA-seq data (Khrameeva *et al.*, 2020) from 33 brain sites of human, chimpanzee, and bonobo were mapped on the masked reference human genome hg38 using the minimap2 (Li, 2016). The hard-masked sequences correspond to the fusion site syntenic regions. Hard-masking was done in order to force unique mapping of the transcripts on the near fusion site region.

A subset of transcripts that were identified on the PhaseDancer assembled subtelomeric sequence extensions: *CBWD2*, *FOXD4L1*, *JMJD7*, *JMJD7-PLA2G4B*, *LINC01881*, *LINC01961*, *MALRD1*, *MAPKBP1*, *PLA2G4B*, *RABL2A*, and *SPTBN5*, was selected to perform the downstream comparative transcriptomic analysis. The selected transcripts coordinates at hg38 genome were downloaded using UCSC hgTables form GENECODE V41 track.

The downstream analysis was performed using a custom-made python script. The analysis starts by defining for each transcript the set of coordinates that describe any of its exomes. For each coordinate, we calculated its coverage using the `pileup` query. Next, for each transcript (for all its exome coordinates) we calculated the average coverage normalised by the sample size (i.e. the total length of all reads in the brain region RNA-seq data in question). The final results were visualised and compared between the brain regions using R-script (Figure 2.8).

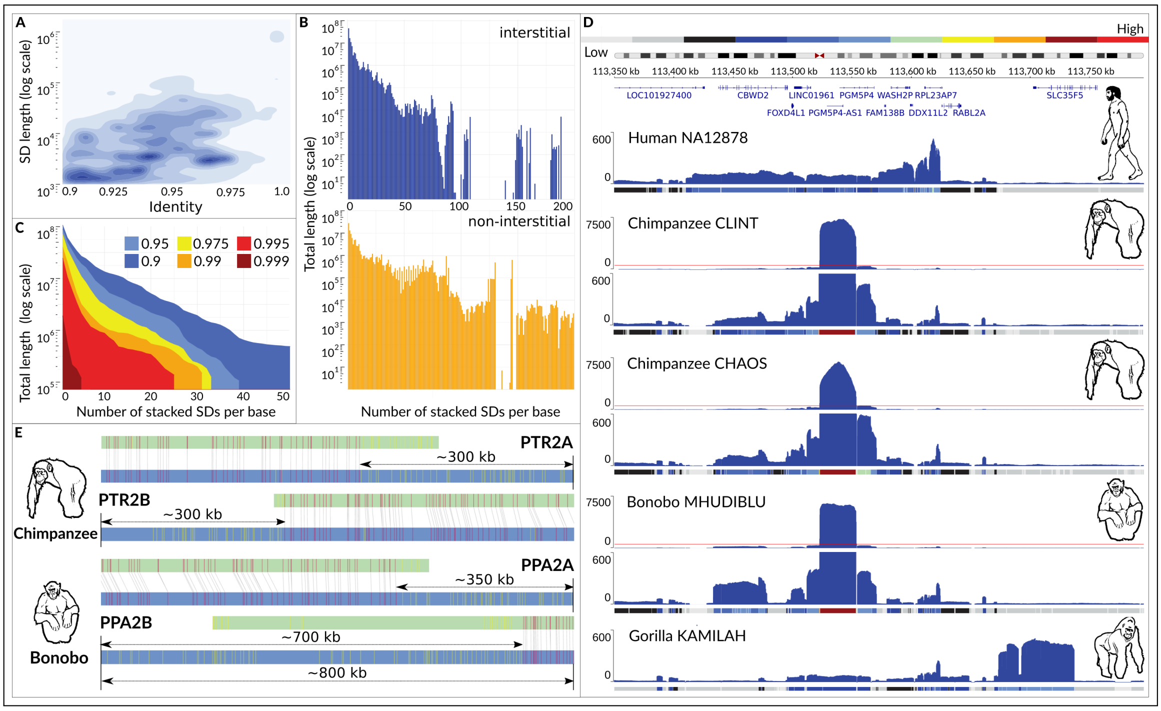


Figure 2.2: An overview of segmental duplications (SDs) characteristics and the study motivation. Based on the most recent T2T human genome assembly: (**A**) A contour plot of the SD abundance given their sequence identity (90–100%, x axis) and the total length (Mb, y-axis, log-scale), where the blue colour intensifies with the increasing number of SDs; (**B**) A barplot of the SDs total length (Mb, log-scale, y-axis) given the total number of SDs copies (x-axis) located at the interstitial (top, blue) and non-interstitial (bottom, yellow) genomic regions; (**C**) An area plot of the SDs' total length (Mb, log-scale, y-axis) for SDs with at least given number of copies (x-axis) and the minimal percent of sequence identity (area colour). Here, the number of stacked SDs per base is the number of reads overlapping a given base position of the reference genome. (**D**) A normalised depth-of-coverage histogram of the aligned whole-genome circular consensus sequencing (CCS) reads in the human (NA12878), two chimpanzees (Clint, Chaos), bonobo (Mhudilbu), and gorilla (Kamilah) genomic regions syntenic to those flanking the HSA2 fusion site. For bonobo and both chimpanzees two depth-of-coverage tracks are shown. The top track presents the full scale of all data, whereas the bottom track zooms-in the coverage of values excluding the extremely high coverage region. The red line on each of the top tracks indicates the y-axis limit of the bottom track. (**E**) Optical genome mapping was used to assess the current incompleteness of the subtelomeric assemblies in chimpanzee and bonobo genomes (panTro5, panTro6, and panPan3). Each of the subtelomeric ends was estimated to lack at least 0.3 Mb of the DNA sequence. Note, the high coverage of the ~31 kb fragment previously found to be amplified about 400 times in the chimp genome (Cheng *et al.*, 2005).

2.3. Results

2.3.1. Design and Implementation of PhaseDancer

In contrast to the existing long-read assemblers that follow the *top-down* paradigm and operating simultaneously on all existing reads, we implemented an approach with contigs generated in a *bottom-up* manner, working with a gradually expanded set of sufficiently similar reads. As a result, our *de novo* assembler can generate several Mb long contigs enriched with SDs.

The algorithm implements an iterative strategy for extending the *initial anchor* sequence by finding the best fitting set of reads to expand the processed *anchor* sequence.

Due to the efficient integration of the state-of-the-art components used in the workflow (see Methods), PhaseDancer generates contigs with the fragments repeated up to several dozens times in the genome with at least 0.1% divergence. The preprocessing time of 200 GB FASTQ data is approximately one hour. The conducted runtime experiments have proven that PhaseDancer is a fast and robust assembler (Fig. 2.5A, B). For example, the targeted assembly of a 1 Mb SD contig (coverage 40x, sequencing error 15%, average read length 18 kb with standard deviation 3 kb) took on average 20 minutes on the server with 56 Intel(R) Xeon(R) E5-2690 v4 @ 2.60GHz CPUs (see Methods).

PhaseDancer is accompanied by two supporting tools, PhaseDancerViewer and PhaseDancerSimulator. PhaseDancerViewer visualises the intermediate results of each algorithm iteration and enables running the assembler in a monitored and semi-supervised fashion, facilitating the PhaseDancer parameters tuning. PhaseDancerSimulator generates *in silico* SD sequences, resulting from various scenarios of a parameter-controlled evolutionary processes. Such synthetic data provide a broad scope of model testing and verification strategies with the *a priori* known dataset.

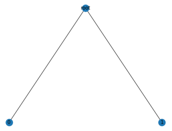
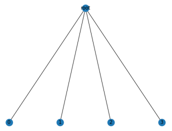
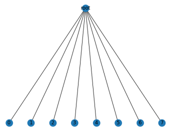
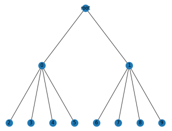
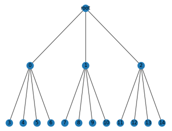
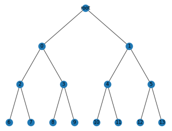
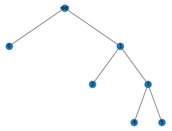
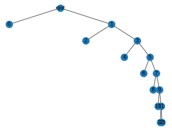
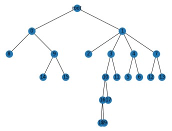
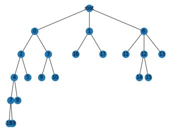
2.3.2. Validation on SD-rich human BAC clone sequences

To validate the PhaseDancer assembly quality, first, we used a set of BAC clones from the haploid CHM13hTERT human cell line (sequenced using PacBio RS II; coverage 45x, N50 20,000), considered as a gold-standard for a validation and benchmarking (Nurk *et al.*, 2022). We employed a validation pipeline commonly used to measure the quality of assemblies on such data, available at `https://github.com/skoren/bacValidation` (Nurk *et al.*, 2020; Shafin *et al.*, 2020). This pipeline evaluates two measures describing the quality of the assembly of the BAC clones sequences: (i) resolving success (BAC clone is considered as resolved if an alignment covers 99.5% of its length), (ii) alignment accuracy (measured as a median of the Phred Quality Scores (Q) (Ewing and Green, 1998) of the alignment identity of the resolved BAC clones). The score Q quantifies the probability (p) of an incorrect base call as $p = 10^{-\frac{Q}{10}}$.

PhaseDancer performance was compared with the results obtained from Flye and Wtdgb2 assemblers that work with the error-prone PacBio reads. Out of 341 BAC clones studied, PhaseDancer resolved 292 clones (85.5%, median Phred Quality Value: 26.81), whereas Flye and Wtdgb2 resolved 91 (26.69%, med. 36.48) and 77 clones (22.58%, med. 30.07), respectively.

Additionally, the impact of the enrichment of BAC clones in SD regions on the assembly results of each tool has been analyzed. To this end, BAC clone sequences were mapped on the T2T genome assembly and using the data from the "SEDEF Segmental Dups" track on UCSC genome browser the average number of stacked SDs for each BAC clone have been computed. The BAC clone was considered as enriched in SDs when the number of SDs is above 0.5. Therefore, 283 of 341 (83 %) BAC clones can be considered as rich in SD

Table 2.1: Assessment of the SDs assembly quality for different tools (columns) in various evolutionary topologies generated by PhaseDancerSimulator (rows). For each table cell: (i) upper value - Phred Quality Score (Q), the larger value the lower error frequency in the assembled sequences; (ii) lower value - a percent of correctly resolved SDs (the expected are sequences from the leaves of the assessed topology). The comparison was evaluated for the following parameters setting of PhaseDancerSimulator: coverage 40x, sequencing error 15%, SD sequence identity 99.5%, average read length 18 kb, read length standard deviation 3 kb, and the simulated SD contig size 0.5 Mb. Timeout - the computation time exceeded 96 hours; N/A - not available, the assembly process failed.

SDs History	PhaseDancer	Canu	Miniasm	Flye	Wtdbg2	SDA
	29.42 100.0%	20.73 50.0%	9.19 50.0%	23.05 50.0%	17.80 50.0%	30.21 0%[1]
	30.19 100.0%	21.64 25.0%	8.92 25.0%	22.91 25.0%	17.65 25.0%	30.41 100%
	30.26 100.0%	20.76 12.5%	8.92 12.5%	22.89 12.5%	- 0.0%	- Timeout
	30.14 100.0%	18.74 12.5%	8.85 12.5%	20.27 12.5%	- 0.0%	- Timeout
	30.08 100.0%	18.71 8.3%	8.76 8.3%	- N/A	- 0.0%	- N/A
	29.83 100.0%	17.34 12.5%	8.53 12.5%	17.901 12.5%	- 0.0%	- Timeout
	30.10 100.0%	19.63 25.0%	8.91 25.0%	21.25 25.0%	17.19 25.0%	- Timeout
	30.24 100.0%	18.25 12.5%	8.45 12.5%	18.22 12.5%	- 0.0%	- Timeout
	30.13 100.0%	18.90 8.3%	8.59 8.3%	- N/A	- 0.0%	- N/A
	30.04 100.0%	17.93 8.3%	8.55 8.3%	- N/A	- 0.0%	- N/A

regions. All assemblers have resolved BAC clones not enriched in SDs. In the resolving BAC clones augmented with SDs PhaseDancer compared favorably against other assemblers. PhaseDancer resolved 81.9%, while Wtdbg and Flye, respectively 7,4% and 12,0% (Fig. 2.3).

Importantly, after backtracking of the PhaseDancer failures, we established that the unresolved BAC clones represented either SD regions with low-coverage or SDs enriched in highly repetitive tandem repeats.

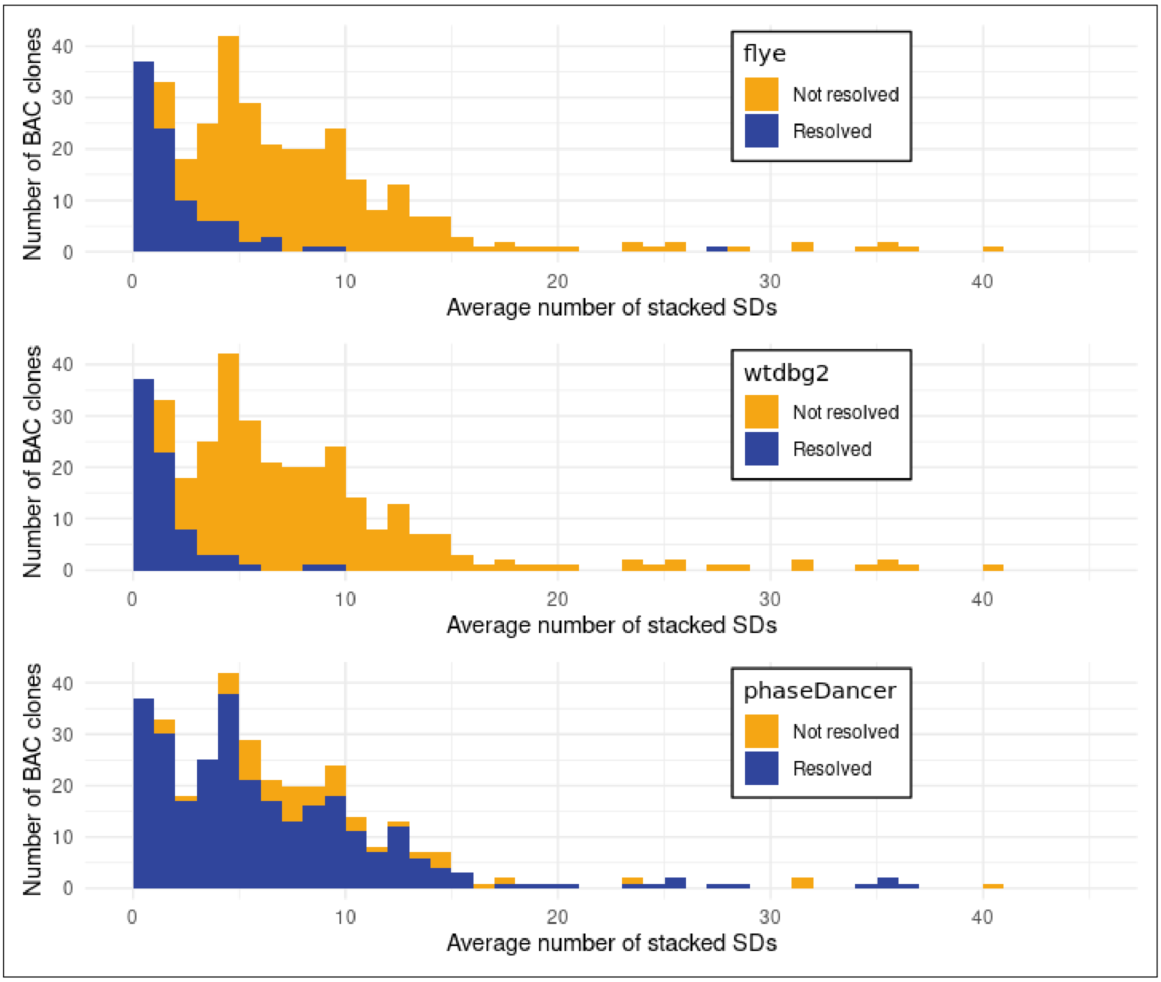


Figure 2.3: The impact of the enrichment of the BAC clones in SDs on the assembly results for Flye, Wtdgb2 and PhaseDancer. Histograms display the number of BAC tasks (y-axis) given the number of stacked segmental duplications within a given BAC sample data (x-axis) for each assembler. The color indicates whether the assembly tasks were resolved (blue) or non-resolved (yellow).

2.3.3. *In silico* verification and benchmarking

To evaluate the accuracy of the PhaseDancer performance, we tested the quality of the assembled sequences from the collapsed SDs generated by PhaseDancerSimulator. We simulated the collapsed SDs using 10 different evolutionary scenarios: flat with two, four, and eight leaves; three types of bifurcating; cascading with four and eight leaves; and two random with 12 leaves (Table 2.1). PhaseDancerSimulator was run with the above-mentioned set of parameters. Additionally, for each of the simulated SDs, random sequences were

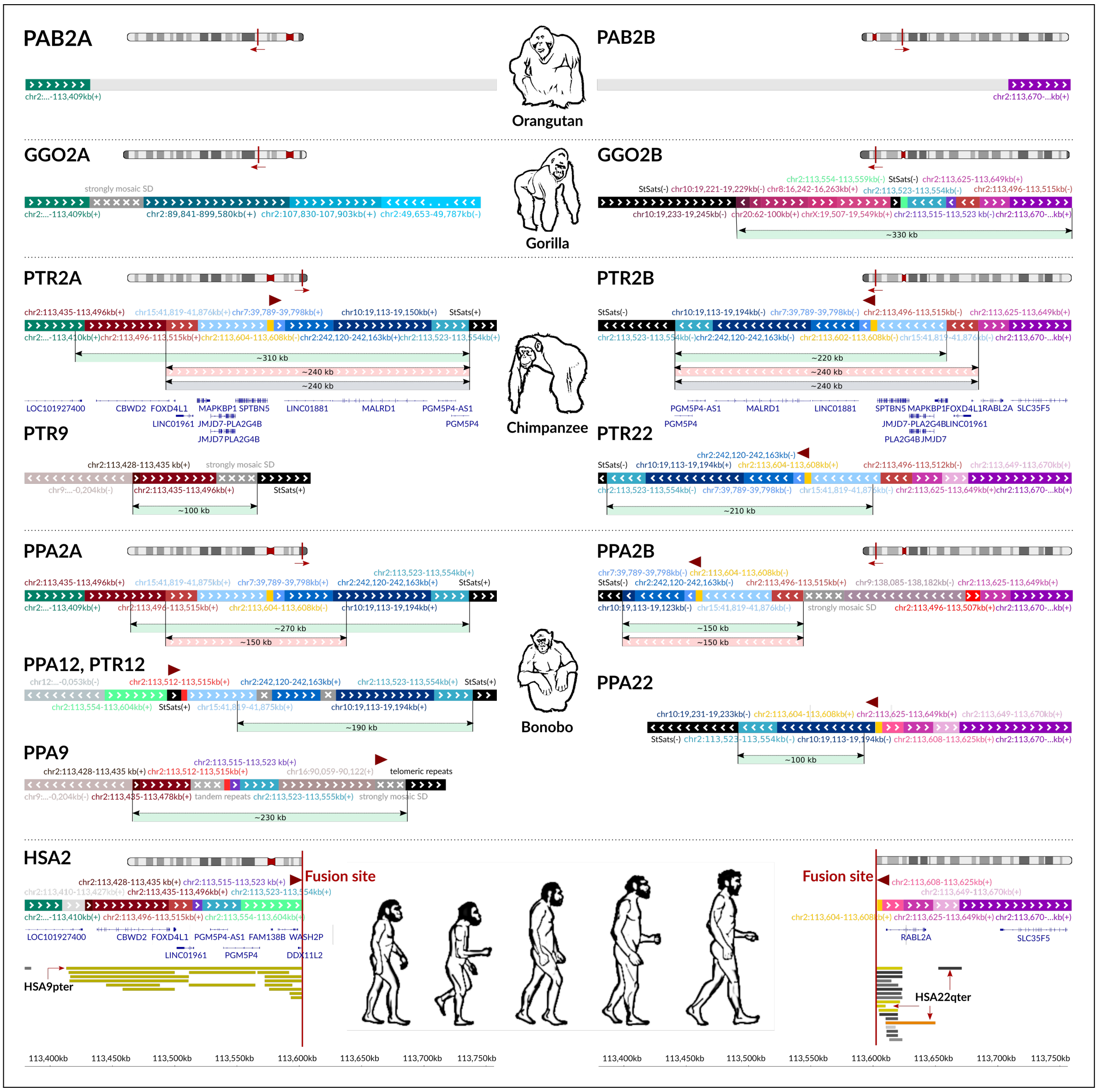


Figure 2.4: Genome architecture flanking the HSA2 fusion site and the syntenic genomic regions in Great Apes and human. From the top, the figure depicts the sequences from: orangutan (PAB) and gorilla (GGO) chromosomes 2Apter and 2Bpter; chimpanzee (PTR) and bonobo (PPA) chromosomes 2Apter, 2Bpter, 9pter, 12pter and 22qter; and human HSA2, all together with the corresponding coding regions track. Each individual contig is represented by a uniquely coloured stripe consistent among species/chromosomes, labelled with the coordinates with respect to the human genome build (hg38) and designated with the arrowheads indicating the DNA strand. Dark grey contigs with white crosses depict strongly mosaic SDs or tandem repeats that cannot be graphically presented in a legible way. Brown arrowheads depict the TAR1 satellite and degenerate telomeric repeats at the HSA2 fusion site and their orthologs in Great Apes. Below each contig assembly a coloured stripe depicts: (i) green - the novel reconstructed assembly along with an approximate size, (ii) pink - the high homology region between chromosomes 2Apter and 2Bpter presumably triggering the fusion event, and (iii) grey - the region that was lost after the fusion event with respect to the HSA2. HSA2 is also equipped with a track of collapsed SDs including ∼190 kb fragment homologous to HSA9pter and three fragments ∼68 kb in size in total homologous to HSA22pter. The azure contig (chr2:113,523-113,554 kb) was found to be amplified ∼400 times in the chimpanzee genome (Cheng *et al.*, 2005).

added at their beginnings and ends. Unique random sequences preceding each collapsed SD portion of the generated sequences were used as an initial anchor sequence for the assembly process.

On such generated synthetic datasets, PhaseDancer was benchmarked against the several commonly used assemblers supporting error-prone NGS long reads: Canu (Koren *et al.*, 2017), Wtdbg2 (Ruan and Li, 2020), Flye (Kolmogorov *et al.*, 2019), Miniasm (Li, 2016), and SDA (Vollger *et al.*, 2019a).

To compare the assembly quality of the above tools, we calculated the Levenshtein distance between all assembled contigs and the simulated SDs. Next, for each assembled contig, we assigned the simulated SD for which: (i) the alignment covers at least 95% of the contig length, and (ii) the alignment Phred Quality Score was highest among all SDs. This assignment procedure allowed us to determine the number of the resolved simulated SDs generated by each assembler.

PhaseDancer has successfully resolved all of the simulated SDs with no alignment of Phred Quality Score lower than 29 (accuracy over 99.874%). Other assemblers managed to resolve at most one reference SD per scenario, and only Canu (Koren *et al.*, 2017), and Miniasm (Li, 2016) produced one sequence for all scenarios. Flye (Kolmogorov *et al.*, 2019) resolved one simulated SD only for models consisting of up to eight SDs, whereas Wtdbg2 (Ruan and Li, 2020) resolved only up to four SDs. Some assemblers failed to complete their assembly task either due to exceeding the 96-hour time limit or execution error during the assembly process (Table 2.1). Additionally, we broadly assessed the PhaseDancer performance on *in silico* reads of various properties provided by PhaseDancerSimulator (Figure 2.5).

2.3.4. Unveiling HSA2 Fusion Event

Following the successful validation of PhaseDancer, we applied our algorithm to the unresolved subtelomeric regions of the selected chromosomes in Great Apes i.e chimpanzee, bonobo, gorilla, and orangutan, syntenic to HSA2, to unravel the mechanism of reduction of the chromosome number during human speciation after divergence from chimpanzee/bonobo. These regions likely reflect high similarity with the ancestral chromosomes 2Apter and 2Bpter, that might have predisposed them for the evolutionary chromosomal fusion event.

Based on classical cytogenetics (Yunis and Prakash, 1982; Turleau *et al.*, 1972; Lejeune *et al.*, 1973; Dutrillaux, 1979) and molecular methods (Ijdo *et al.*, 1991; Kasai *et al.*, 2000; Avarello *et al.*, 1992; Wienberg *et al.*, 1994; Allshire *et al.*, 1988; Wienberg *et al.*, 1990; Wells *et al.*, 1990; Jauch *et al.*, 1992). HSA2 was proposed to have arisen as a product of the end-to-end fusion of telomeric repetitive sequences of the ancestral primate chromosomes 2Apter and 2Bpter. Subsequently, the unstable dicentric chromosome was rescued by a loss of satellite DNA sequences in the vestigial centromere at 2q21.2 (Ijdo *et al.*, 1991; Martin *et al.*, 2002; Miga, 2017; Chiatante *et al.*, 2017; Baldini *et al.*, 1993; Avarello *et al.*, 1992; Lejeune *et al.*, 1973; Dutrillaux, 1979; Roberto *et al.*, 2008). Prior to the fusion, both ancestral chromosomes 2A and 2B underwent ancestral large pericentric inversions, before

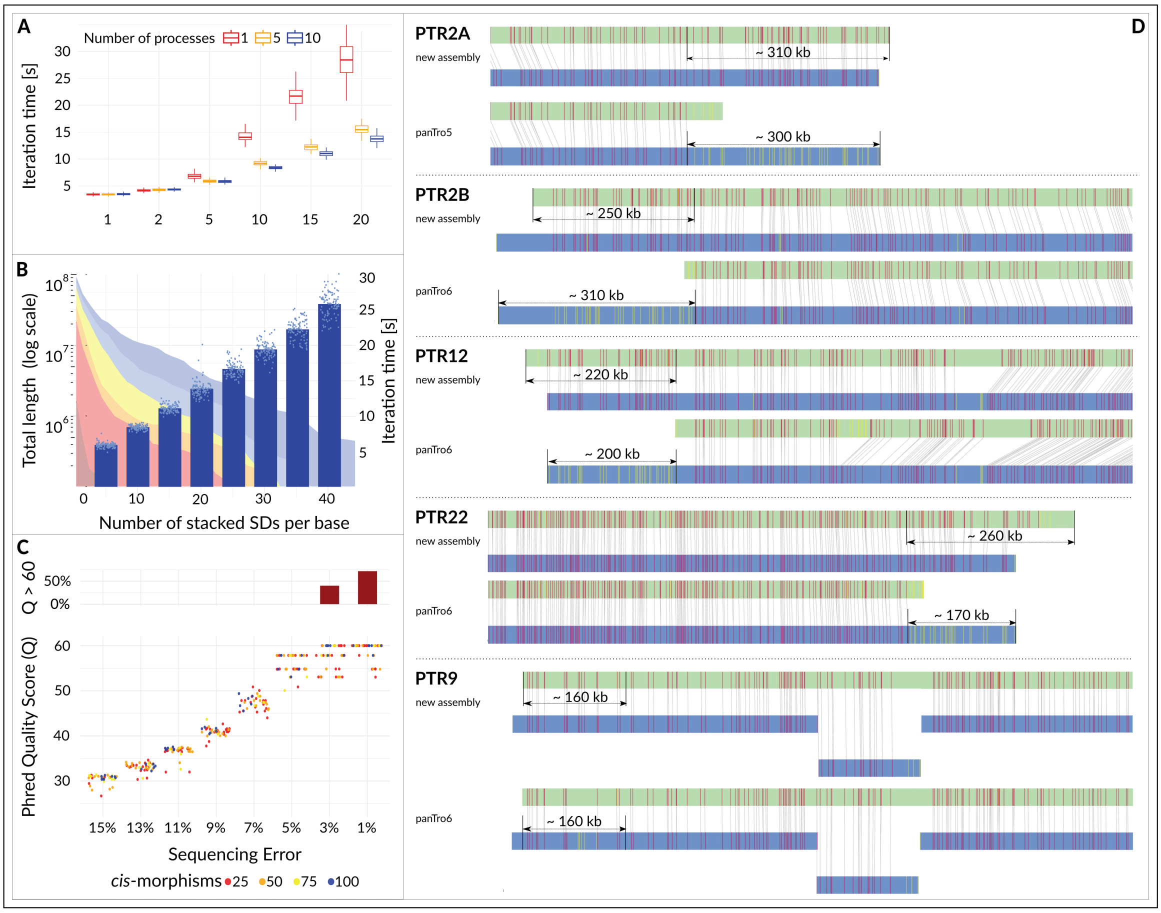


Figure 2.5: Time complexity, feasibility, and correctness of PhaseDancer. (**A**) Computational time performance (y-axis) for different number of stacked SDs (x-axis) and processes (colour scale). Each boxplot represents 100 iterations of PhaseDancer for a given setting. (**B**) Feasibility space for SDs in human. PhaseDancer resolves all SDs with the number of stacked SDs per base as for SDs identified by T2T human genome assembly (area plot, Fig. 2.2C). For a given number of stacked SDs (x-axis) the height of each bar indicates an average runtime of PhaseDancer iteration (right y-axis) along with standard deviation (error bars) and individual measurements (points). (**C**) The evaluation of PhaseDancer assemblies using the Phred Quality Score (Q; y-axis). The samples used for evaluation were generated by PhaseDancerSimulator, with fixed parameters including a coverage of 40x, an average read length of 18 kb, and a read length standard deviation of 3 kb. The x-axis represents different sequencing error levels, while the colour scale indicates different numbers of cis-morphisms per 10 kb window. The additional upper panel in the figure shows the percentage of assembly tasks with no errors (Q > 60) using bar plots. Remarkably, our analysis revealed no significant changes in assembly quality for different PhaseDancerSimulator topologies (SDs evolutionary scenarios). (**D**) Correctness of the PhaseDancer assemblies was assessed using optical genome mapping (OGM). All HSA2 syntenic sites of the chimpanzee genome were in concordance with the corresponding OGM molecules (BssSI enzyme shown).

the chimp-gorilla lineage split (Yunis and Prakash, 1982; Kasai *et al.*, 2000; Wienberg *et al.*, 1994; Roberto *et al.*, 2008) and after the orangutan-gorilla divergence (Yunis and Prakash, 1982; Kasai *et al.*, 2000; Ventura *et al.*, 2011; Wienberg *et al.*, 1994), respectively (Figure 2.6).

We confirmed the incompleteness and partial incorrectness of the latest genome builds of the subtelomeric sequences of Great Apes chromosomes 2Apter and 2Bpter using optical genome mapping (OGM) and direct sequence analysis. The uniqueness and non-recurrence of this event was validated by analysing the human population SNV and SV polymorphisms flanking the HSA2 fusion site (Fig. 2.4, Fig. 2.4 and Tab.2.2).

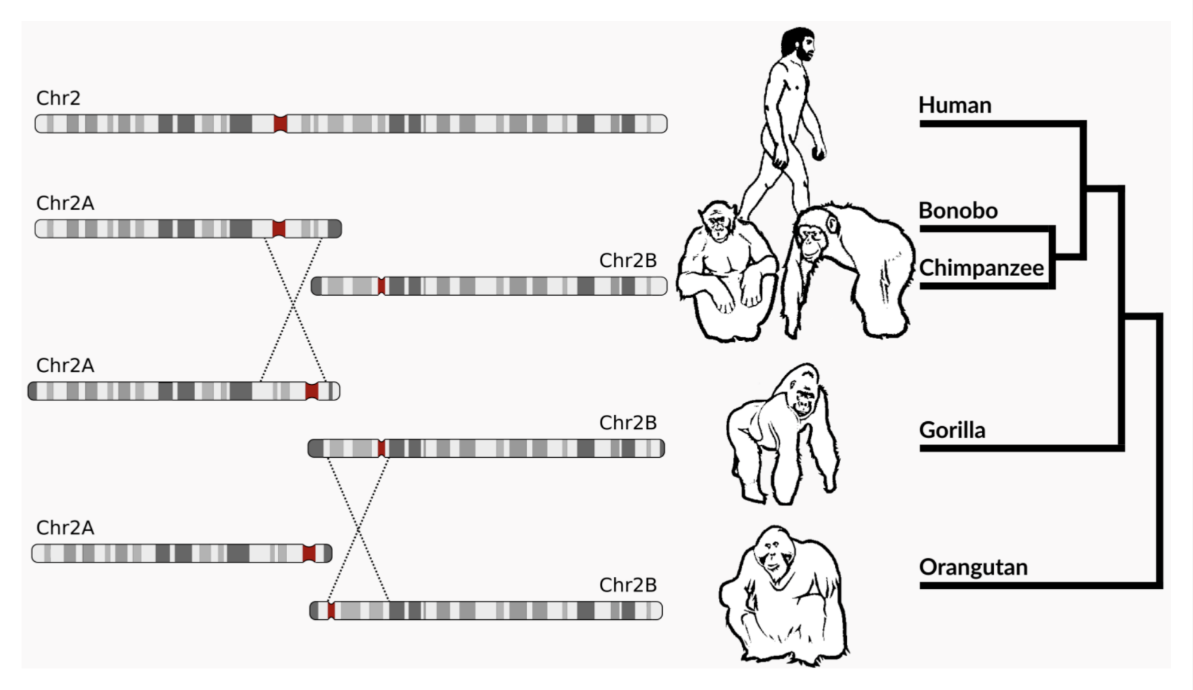


Figure 2.6: Gross inversion events in the course of primate evolution. Note that the orangutan's acrocentric Chr2B was inverted to form the gorilla's Chr2B, and later the chimpanzee and bonobo. Later, a similar event formed the chimpanzee and bonobo metacentric Chr2A after inversion on the gorilla acrocentric Chr2A. Eventually, the chromosomal fusion created the human Chr2 from the ancestral Chr2A and Chr2B and reduced the number of human chromosomes from 48 to 46.

2.3.5. Great Apes genomes analysis

Orangutan. The orangutan genome differs from gorilla genome by one and from chimpanzee, bonobo, and human genomes by two gross chromosomal inversions rearranging them from acrocentric to submetacentric chromosomes. Using PhaseDancer in combination with OGM, we confirmed that the regions syntenic to the HSA2 fusion site map to the latest orangutan genome build showing any structural variations.

Gorilla. PhaseDancer generated assembly extending GGO2Bpter with ~330 kb, reaching the highly repetitive subtelomeric satellites (StSats) regions. The novel fragment of the GGO2Bpter is homologous to the proximal side of the HSA2 fusion. However, inside this fragment, we identified an ~54 kb sequence homologous to the distal side of the HSA2 fusion (chr2:113,496-113,550 kb). The ~44 kb contig on GGO2Apter (Fig. 2.4, the grey contig) is a region that maps to many different locations not related to the fusion site. Using OGM, we confirmed the presence of an erroneously scaffolded ~89 Mb region in the latest GorGor6 assembly.

Chimpanzee. Using OGM, we found false positive breakpoints on PTR2Apter in the latest chimpanzee chromosome build (panTro6) that resulted in placing the subtelomeric region interstitially, whereas no errors were found in the PTR2Bpter subtelomeric region. PhaseDancer extended both PTR2Apter and PTR2Bpter with an ~270 kb sequence, reaching StSats repetitive sequences, each harbouring ~240 kb of the fully homologous fragments. Importantly, the detected homologies encompass a fragment of ~190 kb that was likely deleted during the fusion event, whereas the remaining ~68 kb fragment is homologous to HSA2 near the fusion site. By homology to the human reference genome (including chromosomes 2, 7, 10, and 15), on the deleted fragments we annotated six coding genes: *MAPKBP1*, *JMJD7 PLA2G4B*, *JMJD7-PLA2G4B*, *SPTBN5*, and *MALRD1* and one lncRNA *LINCO1881*. All coding regions were subjected to the downstream transcriptomic analyses and their activity was assessed in different locations of human brain using the RNA-seq transcriptomic data (Khrameeva *et al.*, 2020) (Fig. 2.8).

Interestingly, we found a strong homology between the region chr2:113,554-113,604 kb next to the fusion site and the chimpanzee subtelomeric region at PTR12pter and extended this region towards StSats. As a result, we identified an ~168 kb homology of PTR12pter to both PTR2Apter and PTR2Bpter, adjacent to an ~31 kb fragment that was found to be amplified ~400 times in the chimpanzee genome (Cheng *et al.*, 2005) and homologous also to the region near the HSA2 fusion site (chr2:113,523-113,554 kb; Fig. 2.2D, 2.4). Similarly, sequence homology between the human chromosomal region chr2:113,625-113,670 kb and the chimpanzee subtelomeric region at chromosome 22q led us to explore PTR22qter. The assembled fragment encompasses greater than 240 kb highly homologous fragment between PTR22qter and PTR2Apter and PTR2Bpter, adjacent also to the above-mentioned ~31 kb fragment (as in PTR12pter). Finally, on the PTR9p subtelomeric region, we identified an ~61 kb fragment homologous to the HSA2 fusion site. However, extension towards StSats did not reveal any additional homology to HSA2, PTR2Apter, or PTR2Bpter.

Bonobo. Analogously to the above Great Apes, both PPA2Apter and PPA2Bpter subtelomeric regions were validated using OGM and were extended to the StSats repetitive sequences by ~270 kb and ~120 kb, respectively. Approximately 150 kb of homology was detected between these chromosomes and the fragments of ~190 kb and ~280 kb from PPA2Apter and PPA2Bpter, respectively, were found to be absent on HSA2. Similarly to the chimpanzee genome, because of the discovered homology between HSA2 fusion site and the bonobo chromosomal regions PPA9pter, PPA12pter, and PPA22qter, we assembled their subtelomeric regions revealing strong homologies to PPA2Apter and PPA2Bpter. However, an extension of PPA9pter from the ~61 kb homology region with HSA2 towards StSats confirmed an additional homology (separated by an insertion) to the above-mentioned ~31 kb fragment amplified in chimpanzee (Fig. 2.2D, 2.4) (Cheng *et al.*, 2005). Similarly to chimpanzee genome, the selected transcripts were analysed to determine genes distinguishing the species Fig. 2.8, Methods). Using OGM, we independently validated the presented novel assemblies, extending the current reference genomes of bonobo, chimpanzee, and gorilla, generated using PhaseDancer (data shown for chimpanzee, Fig. 2.5D).

Human. Finally, using PhaseDancer, we assembled the NGS data from ten individuals from

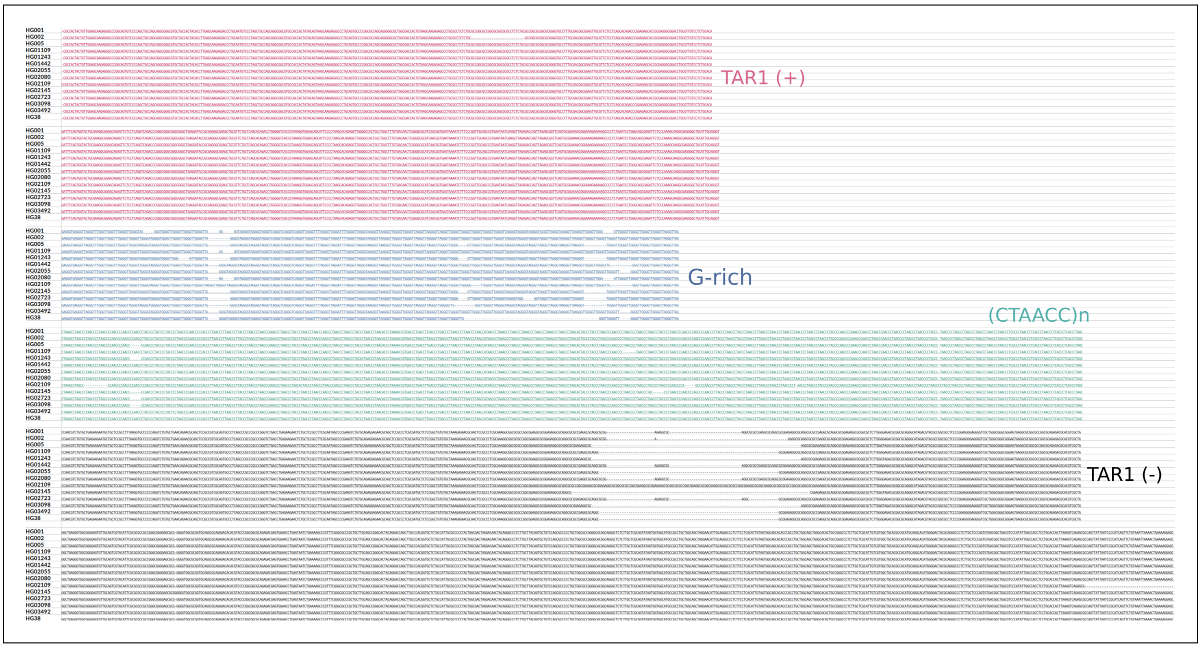


Figure 2.7: Multialignment of the genomic fragments flanking the HSA2 fusion site. Fragments of the assembled contigs flanking the fusion site for 13 human genomes were multialigned by CLUSTALW algorithm with default parameters. For the alignment fragments of the assembled fusion sites contigs were annotated by RepeatMasker (Tab. 2.2. Fragments subsequently annotated as TAR1 satellite, G-rich low-complexity region, (CTAACC)n simple repeat, and inverted TAR1 satellite was used for the multialignment. In the multialignment visualisation, each region is depicted with a distinct colour.

the Human Pangenome Project, T2T Diversity Panel (Wang *et al.*, 2022) and three individuals from the Genome in the Bottle project (Zook *et al.*, 2016). The selected individuals represent five main human superpopulations: African, admixed American, East Asian, European, and South Asian. In particular, we assessed the polymorphisms of the 5 kb region directly flanking the HSA2 fusion site. The selected sequences corresponding to the region were subjected to the downstream analyses using RepeatMasker and multialigned to detect any possible genomic variety. No significant structural variations (i.e. duplications, deletions, inversions, indels) were detected (Tab. 2.4) and Fig. 2.7).

2.3.6. Analyses of the newly assembled two chimpanzee genomes

To confirm the structure of the assembled genomic extensions obtained using PhaseDancer, we incorporated additional NGS long-reads from two different chimpanzee individuals sequenced for this study (Chaos and Toby). The datasets are publicly available in NCBI SRA repositories under the accession number PRJNA905805 (Methods). Our analyses of the WGS data confirmed the computed subtelomeric structures, and found no significant polymorphisms (data not shown), further confirming the structure of the obtained assemblies.

Table 2.2: RepeatMasker analysis of HSA2 fusion site flanking regions of 3 human genomes from Genome in the Bottle project repository and 10 human genomes from T2T Diversity Panel. Contigs spanning the fusion site were assembled using PhaseDancer for 3 human genomes from Genome in the Bottle project repository: HG001, HG002, HG005 and for 10 human genomes from T2T Diversity Panel: HG01109, HG01243, HG02080, HG03098, HG02055, HG03492, HG02723, HG02109, HG01442, HG02145, Each of these contigs was aligned to the 5kb region directly flanking the HSA2 fusion site (chr2:113,601,000-113,605,999 (hg38)). The table lists the output of the RepeatMasker for the aligned fragments for the repeats immediately flanking the fusion site (TAR1 satellite, G-rich low-complexity region, (CTAACC)n simple repeat, and inverted TAR1 satellite).

score	% div.	% del.	% ins.	query begin	query end	strand	repeat	class/family	repeat begin	repeat end
Reference hg38										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
136	13,6	0,4	2,5	1683	1929	+	G-rich	Low_complexity	1	242
329	8,5	1,5	3,5	1930	2478	+	(CTAACC)n	Simple_repeat	1	538
5395	10,3	2,3	1,3	2479	3531	C	TAR1	Satellite/subtelo	(316)	1063
HG001										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
173	13,4	0,3	2,3	1683	1990	+	G-rich	Low_complexity	1	302
329	8,5	1,5	3,5	1991	2539	+	(CTAACC)n	Simple_repeat	1	538
5489	10,2	2,4	1,6	2540	3624	C	TAR1	Satellite/subtelo	(316)	1092
HG002										
2701	12	0,8	0,9	976	1652	+	TAR1	Satellite/subtelo	1	676
181	13,3	0,3	2,2	1654	1973	+	G-rich	Low_complexity	1	314
329	8,5	1,5	3,5	1974	2522	+	(CTAACC)n	Simple_repeat	1	538
5395	10,3	2,3	1,3	2523	3575	C	TAR1	Satellite/subtelo	(316)	1063
HG005										
2863	11,8	0,7	1,3	977	1682	+	TAR1	Satellite/subtelo	1	702
165	13,8	0,3	2,7	1684	1986	+	G-rich	Low_complexity	1	296
303	7,3	1,8	3,9	1987	2532	+	(CTAACC)n	Simple_repeat	1	535
5200	10,7	4,1	0,6	2533	3566	C	TAR1	Satellite/subtelo	(319)	1069
HG01109										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
181	13,3	0,3	2,2	1683	2002	+	G-rich	Low_complexity	1	314
325	8,6	1,5	3,4	2003	2544	+	(CTAACC)n	Simple_repeat	1	532
5395	10,3	2,3	1,3	2545	3597	C	TAR1	Satellite/subtelo	(316)	1063
HG01243										
2872	11,6	0,7	1,3	977	1682	+	TAR1	Satellite/subtelo	1	702
161	13,6	0,3	2,8	1684	1980	+	G-rich	Low_complexity	1	290
325	8,4	1,5	3,6	1981	2523	+	(CTAACC)n	Simple_repeat	1	532
5200	10,7	4,1	0,6	2524	3560	C	TAR1	Satellite/subtelo	(316)	1072
HG01442										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
176	13,6	0,3	2,3	1683	1996	+	G-rich	Low_complexity	1	308
329	8,5	1,5	3,5	1997	2545	+	(CTAACC)n	Simple_repeat	1	538
5489	10,2	2,4	1,6	2546	3630	C	TAR1	Satellite/subtelo	(316)	1092
HG02055										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
179	13,7	0,3	2,2	1683	2002	+	G-rich	Low_complexity	1	314
329	8,5	1,5	3,5	2003	2551	+	(CTAACC)n	Simple_repeat	1	538
5395	10,3	2,3	1,3	2552	3604	C	TAR1	Satellite/subtelo	(316)	1063
HG02080										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
180	12,8	0,3	2,3	1683	1996	+	G-rich	Low_complexity	1	308
329	8,5	1,5	3,5	1997	2545	+	(CTAACC)n	Simple_repeat	1	538
5473	10,3	2,4	1,6	2546	3630	C	TAR1	Satellite/subtelo	(316)	1092
HG02109										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
180	12,8	0,3	2,3	1683	1996	+	G-rich	Low_complexity	1	308
329	8,5	1,5	3,5	1997	2545	+	(CTAACC)n	Simple_repeat	1	538
5473	10,3	2,4	1,6	2546	3630	C	TAR1	Satellite/subtelo	(316)	1092
HG02145										
2872	11,6	0,7	1,3	977	1682	+	TAR1	Satellite/subtelo	1	702
165	13,8	0,3	2,7	1684	1986	+	G-rich	Low_complexity	1	296
322	8,3	1,5	3,6	1987	2523	+	(CTAACC)n	Simple_repeat	1	526
5195	10,9	2,4	1	2524	3545	C	TAR1	Satellite/subtelo	(316)	1036
HG02723										
2872	11,6	0,7	1,3	977	1682	+	TAR1	Satellite/subtelo	1	702
159	14,1	0,3	2,8	1684	1980	+	G-rich	Low_complexity	1	290
303	7,3	1,8	3,9	1981	2526	+	(CTAACC)n	Simple_repeat	1	535
5248	10,6	4,1	1,1	2527	3592	C	TAR1	Satellite/subtelo	(319)	1098
HG03098										
2872	11,6	0,7	1,3	977	1682	+	TAR1	Satellite/subtelo	1	702
161	13,6	0,3	2,8	1684	1980	+	G-rich	Low_complexity	1	290
325	8,4	1,5	3,6	1981	2523	+	(CTAACC)n	Simple_repeat	1	532
5200	10,7	4,1	0,6	2524	3560	C	TAR1	Satellite/subtelo	(316)	1072
HG03492										
2872	11,6	0,7	1,3	976	1681	+	TAR1	Satellite/subtelo	1	702
181	13,3	0,3	2,2	1683	2002	+	G-rich	Low_complexity	1	314
326	8,7	1,5	3,5	2003	2551	+	(CTAACC)n	Simple_repeat	1	538

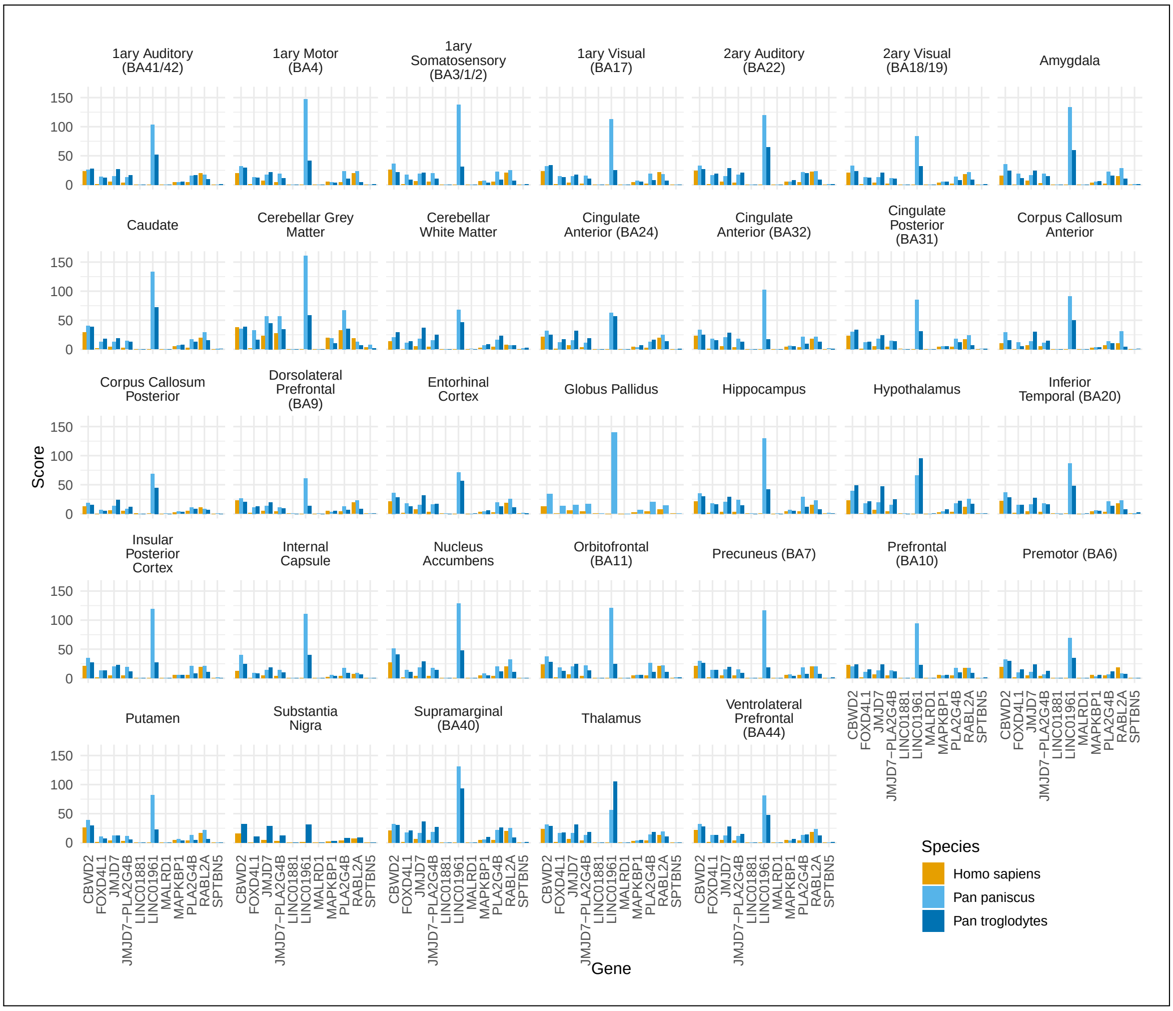


Figure 2.8: Expression levels of 11 transcripts in chimpanzee, bonobo, and human (*CBWD2*, *FOXD4L1*, *JMJD7*, *JMJD7-PLA2G4B*, *LINC01881*, *LINC01961*, *MALRD1*, *MAPKBP1*, *PLA2G4B*, *RABL2A*, and *SPTBN5*) found on the extensions of the subtelomeric regions assembled with PhaseDancer. No data available for: Substantia Nigra (Pan paniscus), Globus Pallidus (Pan troglodytes).

2.4. Discussion

We have shown the extent to which PhaseDancer can serve as an efficient, robust, and reliable tool resolving complex SD-rich genomic regions. Compared to the latest, commonly used assemblers, it provides the most accurate data, even for SDs with highly complex structures in the shortest time. Moreover, such tasks are accomplished also for the error-prone long reads.

Consequently, PhaseDancer has enabled substantial and robust extensions of the Great Apes subtelomeric regions evolutionarily important for the HSA2 formation. We have provided the validated and publicly available tool relying on the currently most efficient software and technologies that can be further developed and extended also at the community-based level.

The results of our assemblies have allowed us to propose a scenario of the evolutionary formation of the HSA2 fusion involving not only chromosomes 2Apter and 2Bpter as

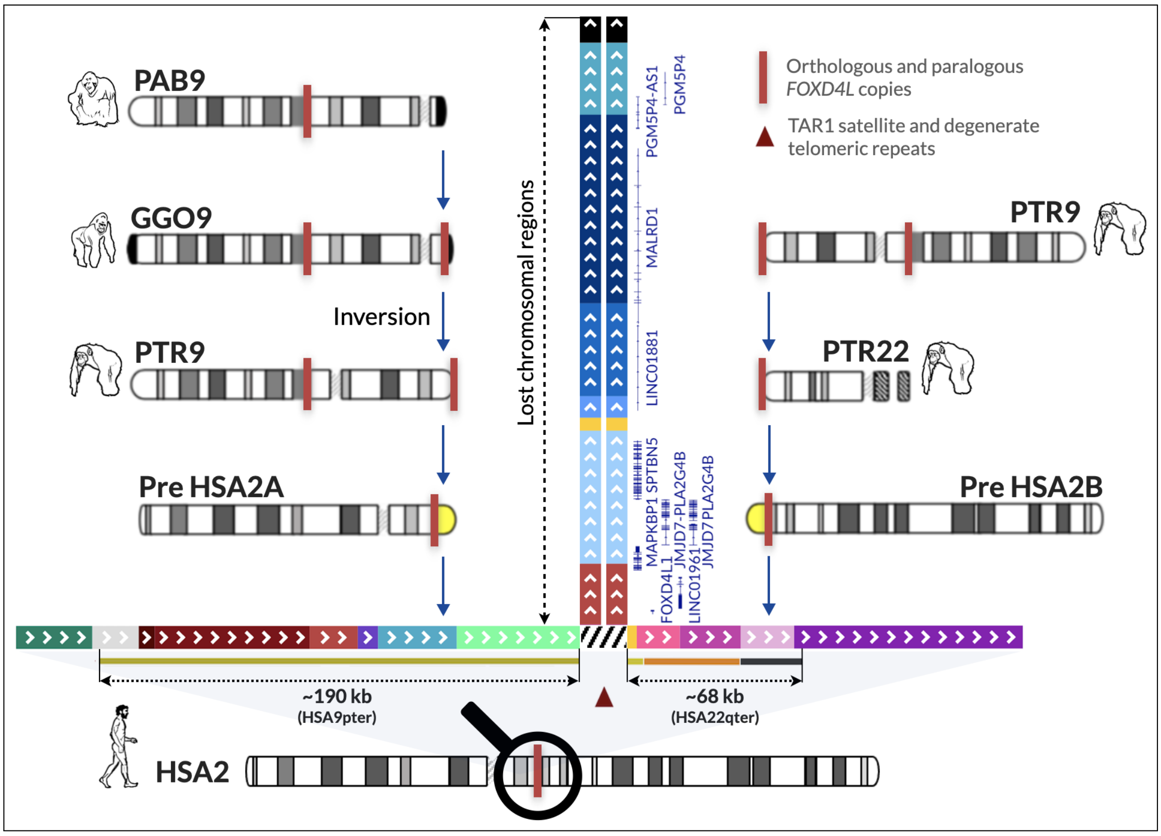


Figure 2.9: The proposed model for the evolutionary HSA2 fusion event based on the assembled SD-rich subtelomeric sequences in Great Apes chromosomes, absent in the reference genomes. The fusion site is flanked proximally and distally, respectively, by the ~190 kb and ~68 kb SDs homologous to human chromosomes 9p24.3 and 22q13.33 (98.9% and 97.8-99.1% sequence identity). The ~190 kb fragment harbouring *FOXD4L1* (red solid rectangle) (Fig. 2.4), and likely originating from an ancestral locus syntenic to chromosome 9q21.11 in human, was previously shown to be duplicatively transposed to chromosome PTR2Apter after gorilla had branched off the common chimp-human ancestor lineage (Martin *et al.*, 2002; Ventura *et al.*, 2012; Lese *et al.*, 1999; Wong *et al.*, 2004). Both copies flank the evolutionarily pericentromeric inversion in the human and chimp genomes that arose after the gorilla divergence (Martin *et al.*, 2002; Fan *et al.*, 2002a; Wong *et al.*, 2004). We have proposed that a portion of the PTR9pter copy was also copied onto chromosome PTR22qter and later PTR2Bter before the gorilla-chimp divergence (Martin *et al.*, 2002; Fan *et al.*, 2002a; Ning *et al.*, 1996; Wong *et al.*, 1999). Importantly, our assemblies revealed substantially long homology (~190kb) between the lost fragments (within the yellow band) of the ancestral chromosomes 2Apter (Pre HSA2A) and 2Bpter (Pre HSA2B) that might have served as a substrate of misalignment during meiosis. The fusion occurred within TAR1 satellite and degenerate telomeric repeats present in both Pre HSA2Apter and Pre HSA2Bpter. Submicroscopic subtelomeric rearrangements in human are relatively common cause of genomic imbalances in patients with developments delay/intellectual disability (Flint *et al.*, 1995). Analyses of these sequences showed that two copies of the following six protein coding genes *FOXD4L1*, *JMJD7-PLA2G4B*, *MAPKBP1*, *PGM5P4*, *SPTBN5*, *CBWD2*, and *MALRD1* and three lncRNAs, *LINC01881* and *LINC01961*, and *PGM5P4-AS1* might have been lost during the fusion event (Fig. 2.4, Fig.2.8).

hypothesised in the current models, but also 9pter and 22qter chromosomes (Fig. 2.9). The existing reference genome sequences of the SD- and StSat-rich subtelomeric regions in the majority of Great Apes chromosomes remain, to a large extent, incomplete. Corroboratively, our assembled sequences of chromosomes 2Apter and 2Bpter in chimp and gorilla are in concordance with the previously published results of the FISH studies with the human cosmid and fosmid probes from the HSA2 fusion site (Fan *et al.*, 2002a; Ventura *et al.*, 2012).

Supporting the notion of Ventura *et al.* (Ventura *et al.*, 2012) that the pericentric inversion of chromosome 2A predisposed the chimpanzee and human genomes to formation of

StSat-rich subtelomeric heterochromatin, whereas the HSA2 fusion prevented our genome from these expansions, we found multiple copies of two unstable genomic segments admixed with the StSat repetitive DNA sequences on the subtelomeric regions of chromosomes 2Apter and 2Bpter in chimpanzee and bonobo as well as on chromosome 2Bpter in gorilla. The copies of the above-mentioned ~31 kb fragment mapping proximally to the fusion site and amplified ~400 times in the chimp genome (Cheng *et al.*, 2005) are admixed to StSat sequences on chromosomes 2Bpter in chimp and 2Apter in bonobo (Fig. 2.2D, 2.4, 2.9). Moreover, the copies of the ~82 kb block 1 (chr10:19,112,612-19,194,164) and the ~43 kb block 2 (chr10:19,238,586-19,281,823) (Marques-Bonet *et al.*, 2009a; Ventura *et al.*, 2012), originating from the ancestral locus orthologous to HSA 10p12.31 and expanded in gorilla with greater than 100 copies and 23-50 copies in chimpanzee and bonobo, but present only in a single copy in human (Fig. 2.10), are directly admixed to StSat sequences on chromosome 2Bpter both in bonobo (chr10:19,112,645-19,123,078) and in gorilla (chr10:19,220,718-19,229,071 and chr10:19,233,190-19,244,822) as proposed by Ventura *et al.* (Ventura *et al.*, 2012).

Out of six protein coding genes (each in two copies) *FOXD4L1*, *JMJD7-PLA2G4B*, *MAPKBP1*, *PGM5P4*, *SPTBN5*, *CBWD2*, and *MALRD1* and three lncRNAs *LINC01881* and *LINC01961*, and *PGM5P4-AS1* that might have been deleted during the fusion event (Fig. 2.4, 2.9), thus far, only *MAPKBP1* has been disease-related in human in an autosomal recessive manner (MIM 617271). Interestingly, *FOXD4*, a member of the forkhead/winged helix-box transcription factor gene family, highly conserved among vertebrates, has been shown recently to play an important role in brain development. In *Xenopus* embryo, *Foxd4l1.1* (previously *Foxd5a/b*), known to play an essential role in maintaining an immature neural fate by regulating several neural transcription factors (Yan *et al.*, 2009, 2010), was found to strongly inhibit mesoderm- and ectoderm-specific marker genes to maintain neural fate by negatively regulating Chordin transcription (Kumar *et al.*, 2021b). In mice, *Foxd4*, required in the transition of the mESCs from pluripotency to neuroectoderm precursor cells, was found to be essential in the anterior mesoderm and in the anterior neuroectoderm for rostral neural tube closure and neural crest specification during head development. Interestingly, loss of *Foxd4* manifested with craniofacial malformations and neural tube closure defects (McMahon *et al.*, 2021). *Foxd4* in mice is also essential for establishing neural cell fate and for neuronal differentiation (Sherman *et al.*, 2017). Loss of *FOXD4* in human was proposed to be responsible for developmental delay in patients with Chromosome 9p deletion (9p-) syndrome (MIM 158170) (Ng *et al.*, 2020). However, the *FOXD4* gene paralogs have not been disease associated, likely because of their multi-copy redundancy. Of note, we found increased expression of all human *FOXD4* paralogs in cerebellum and *FOXD4L2* in tibial nerve (https://gtexportal.org/), suggesting their potential role in bipedalism.

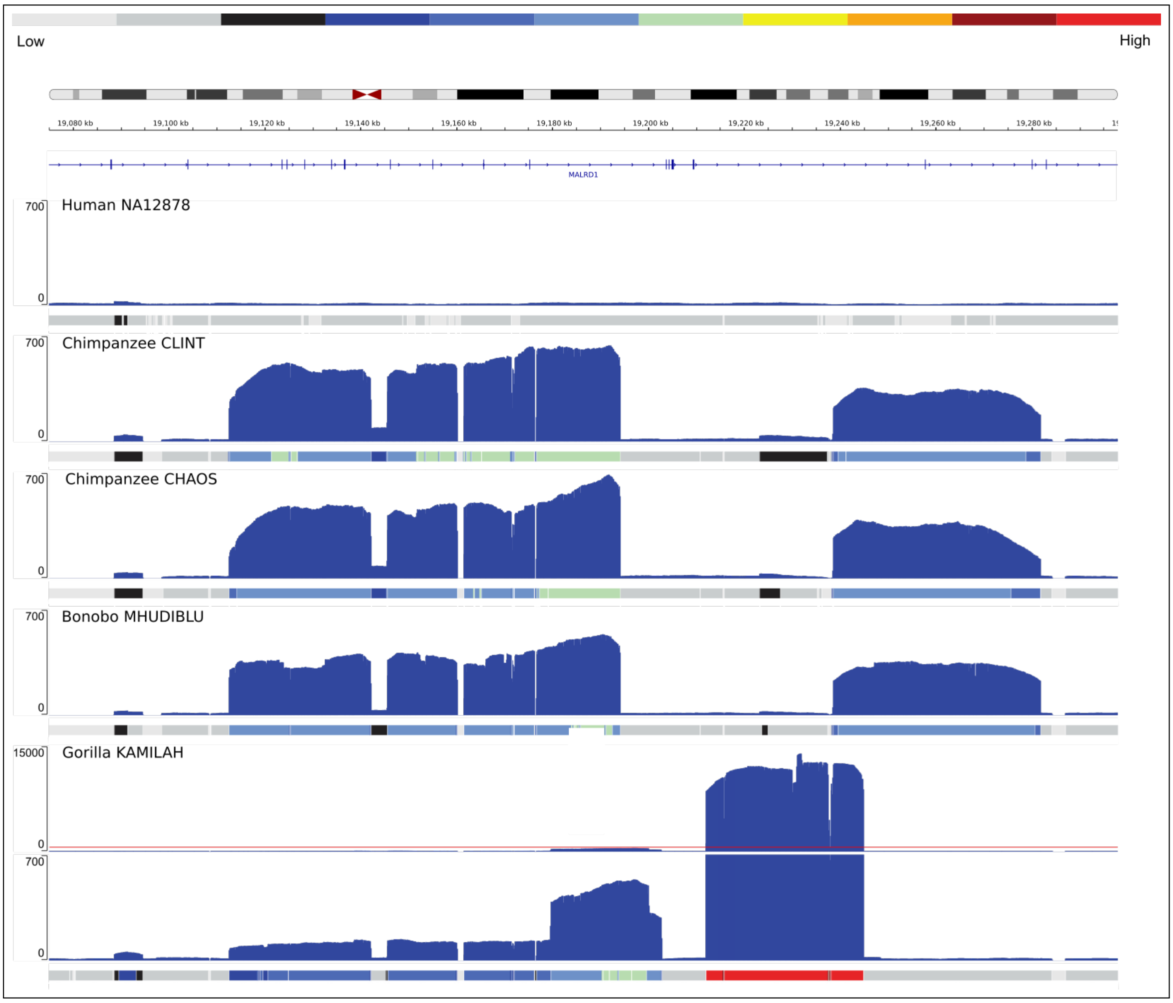


Figure 2.10: Normalised depth-of-coverage histogram of the aligned whole-genome CCS reads of a 225-kbp region of human chromosome 10 (chr10:19075000-19300000, NCBI hg38) in human (NA12878), two chimpanzees (Clint, Chaos), bonobo (Mhudilbu) and gorilla (Kamilah). This region is segmentally duplicated in the chimpanzee, bonobo and gorilla mainly in subtelomeres. In gorilla, two depth-of-coverage tracks are shown. The Y-axis limit of the top track allows for the presentation of all data. The Y-axis limit of the bottom track allows for the presentation of values apart from the region with extremely high coverage. Red line on the top track marks the Y-axis limit of the bottom track.

HSA2 was estimated to have occurred 0.74 Mya (Dreszer *et al.*, 2007), ~3.5 Mya (Miga, 2017), greater than 4 Mya (Ventura *et al.*, 2012), between 1-6 Mya (Fan *et al.*, 2002a), and between 5-6 Mya (Chiatante *et al.*, 2017). Most recently, by re-analysing the enrichment of weak-to-strong (AT to GC) substitutions around the fusion site, we dated its formation at ~0.9 Mya with an upper boundary of ~1.5 Mya (Poszewiecka *et al.*, 2022a). However, it is tempting to speculate that HSA2 fusion was a major evolutionarily event that had initiated the separation of *Hominina* from *Pan* (chimpanzee and bonobo) and introduced the reproductive barrier between them. Moreover, the early HSA2 stabilisation by fusion of chromosomes 2A and 2B harbouring these genome destabilising chr2 and chr10 segments could explain the absence of the StSat-rich cap sequences (StSat, SatIII, and rDNA) expanded in gorilla, chimpanzee, and bonobo (Ventura *et al.*, 2012). Our genomic analyses in 13 individuals revealed no evidence of variability at the HSA2 fusion site, including the terminal degenerate repeats as well as the flanking complex SDs in humans (Fig. 2.7, Tab. 2.2), implying that HSA2 fusion was most likely a nonrecurrent event. We have proposed that large paralogous sequences on distal chromosomes 2Apter and 2Bpter, representing, respectively, orthologous regions on chromosomes 9pter and 22qter in Great Apes, might have facilitated meiotic misalignment between these chromosomes. Our computational analyses of the Great Apes genomes revealed that the ~800 bp TAR1 satellite and degenerate telomeric repeats present at the HSA2 junction site have orthologous copies in both PTR2Apter and PTR2Bpter, indicating where the break and fusion might have occurred (Fig. 2.9).

PhaseDancer is a cutting-edge tool for targeted *de-novo* genomic assemblies, including complex SD-rich regions. The potential applications include also: (i) assembly of the subtelomeric and complex regions of human chromosomes, (ii) fast assembly of the unique genomic regions, and (iii) assessment of the SD copy-number. In addition to the presented evolutionary events it also has potential in personalised medicine for targeting patient-specific SD-related disorders.

2.5. Conclusions and Further Research

In this chapter an innovative assembler PhaseDancer specialized at resolving SDs has been presented. Its utility has been proven by successful assembly of subtelomeric regions of selected chromosomes of primates and validation on real and synthetic data.

Since the architecture of PhaseDancer is modular, improvements of individual moving parts responsible for different assembly steps are seamless. Useful advancement may be introduced in the clustering module by enhancing the cis-morphisms detection procedure, and in the method of grouping reads into clusters. Another room for improvement is the assembly step, where currently third-party general-purpose assembler is used. This tool can be replaced by a custom multi-alignment procedure utilizing the information about the location of mappings onto the anchor sequence, which is now neglected. Inclusion of this information will definitely improve the quality of the assembly in regions enriched in short tandem duplication, which currently used assembler, wtdbg2, tends to collapse.

Moreover, future improvements may take advantage of other types of sequencing data in the assembly process. The short-read data can be used in two procedures: cis-morphisms detection and polishing of the assembly results. However, it is worth nothing that all of these changes have to be done with great attention, as they may introduce errors, especially when an assembled fragment is very similar to other fragments from the genome. Optical genome mapping data also can be incorporated into the workflow in the cluster selection step.

The idea of local assembly using long-read sequencing technology can be further developed and applied. Possible use case is the assembly of clinically relevant fragments of genomes, which can be subsequently visually inspected using the accompanied PhaseDancerViewer. These include assembly of complex chromosomal rearrangements like duplication–triplication/inversion–duplication (DUP–TRP/INV–DUP) syndromes (Carvalho and Lupski, 2016; Schuy *et al.*, 2022). PhaseDancer may also be utilized for closing gaps in the existing assemblies, which typically are enriched with SDs. Another application is the extension of the assembly results obtained from very fast assemblers, but which are "relatively conservative in (segmentally) duplicated regions", like Shasta (Shafin *et al.*, 2020).

3
Revised time estimation of the ancestral human chromosome 2 fusion

"A czy przyroda kolebka
Myślała kiedyś dokładnie
Po co jej wielkie mamuty
Ani wygląda to ładnie
Ani z nich skóra na buty.
Nie ma co pytać, koledzy: robiła i tak jej wyszło.
Nikt nie wymyślał specjalnie tego w czym żyć nam przyszło.
Uprzedzam o tym lojalnie."

— Jacek Kleyff, "Huśtawki"

The reduction of the chromosome number from 48 in the Great Apes to 46 in modern humans, as discussed and modelled in the previous chapter, is thought to be the result of the end-to-end fusion of two ancestral non-human primate chromosomes, forming the human chromosome 2 (HSA2). Genomic signatures of this event includes the presence of inverted telomeric repeats at the HSA2 fusion site, which was extensively analysed in genomes of selected individuals representing five main human superpopulations in Chapter 2. Additionally, remnants of the ancestral centromere are marked by a block of degenerate satellite sequences. It has been estimated that this fusion arose between 0.74 and 4.5 million years ago (Mya).

The ancestral chromosomal fusion, creating the human chromosome 2 (HSA2) and reduction of the chromosome number from 48 in the Great Apes to 46 in modern humans was described nearly four decades ago (Yunis and Prakash, 1982; Ijdo *et al.*, 1991; Luke and Verma, 1995). To better understand this event, the 2q13–q14.1 fusion site has been analyzed using different computational and molecular methods.

Fluorescence in-situ hybridization (FISH) analyses confirmed that two ancestral Great Ape chromosomes fused at their telomeric repeats to form the HSA2 (Kasai *et al.*, 2000). Subsequent studies confirmed also the presence of multiple subtelomeric duplications (SD) with other autosomal chromosomes (Hillier *et al.*, 2005) and described the gene content at the fusion site (Fan *et al.*, 2002b,a). Additionally, the comparison of SDs between the chimpanzee and human genomes not only enabled estimation of the genomic duplication rate, but also suggested SDs as the key cause of transcriptional differences between species and the formation of the ancestral fusion. A 40 kb SD near the fusion site has been identified in 300-500 copies in the chimpanzee genome but only in 4-5 copies in the modern human genome (Cheng *et al.*, 2005).

Using the yeast genome with the functional single-chromosome as a model, it was shown that the reduction of the number of chromosomes does not always have to lead to fatal genetic dysfunctions (Luo *et al.*, 2018; Shao *et al.*, 2018).

These genomic observations have raised questions about the time scope when this gross chromosomal aberration arose. Dreszer *et al.* (2007) have proposed a time estimation method based on the analysis of the fixed substitutions in the human and chimpanzee genomes since their divergence from the common ancestor. The authors have referred to the biased gene conversions (BGCs) occurring due to the mutagenic recombination events (Strathern *et al.*, 1995) and the associated DNA repair processes to favor strong (GC) versus weak (AT) nucleotide pairs at the non-Watson-Crick heterozygous sites in heteroduplex DNA (Meunier and Duret, 2004). Importantly, it has been broadly discussed that BGC may be one of the main evolutionary mechanism (Marais, 2003; Duret and Galtier, 2009). However, Dreszer *et al.* (2007) observed that particularly weak-to-strong (AT to GC) substitutions over-represented locally, e.g. clustering densely near the telomeres of the autosomal chromosomes. Furthermore, using the Unexpected Bias Clustered Substitutions (UBCS) statistics measuring the bias towards weak-to-strong substitutions among the clustered substitutions, a similar over-representation for human and chimpanzee orthologous regions was detected. This observation suggested the existence of a stable evolutionary force that had led to the formation of the biased clusters of substitutions. As expected, around the ancestral HSA2 fusion site, an additional local maximum of the UBCS statistic values was determined. To approximate the time of the fusion event Dreszer *et al.* (2007) assumed that: (i) human-chimpanzee split had occurred 6 Mya and (ii) the rate of the UBCS accumulation is constant. Based on that, they compared the reduction of the bias in the regions near the fusion site with the orthologous telomeric sites of the chimpanzee chromosomes 2a and 2b. As a result, they estimated the fusion time at 0.74 Mya with a 95% confidence interval 0–2.81 Mya.

A phylogenetic analysis of the SVA elements (i.e. composite repetitive elements named after its main components, SINEs, VNTRs and *Alu*s) was performed by Wang *et al.* (2005). The authors showed that within this hominid-specific retroposone family, both SVA-E and the SVA-F subfamilies are restricted to the human lineage. Additionally, based upon the nucleotide divergence, they estimated the expansion time of these subfamilies at 3.5 Mya (with a GC content-dependent range of 2.5–4.5 Mya), which provided a lower bound of the human-chimpanzee speciation event.

In support of these estimations, using the next generation sequencing (NGS) with a high read coverage, Meyer *et al.* (2012) have reconstructed a genome of the *Denisovans*, an extinct relative of the *Neandertals*, and identified an evidence of the HSA2 fusion event. These findings corroborated the theory that the *Denisovans* (and presumably also the *Neandertals*) had shared the fused HSA2 with modern humans. Moreover, the studies on the shared centromere sequence organization in the *Denisovan* and *Neandertal* genomes provided an additional premise that the HSA2 fusion arose prior to our last common ancestor with *Hominins* (Miga, 2017).

We present the revised estimation of the HSA2 fusion time. Our results are twofold. First, we developed a novel algorithm for the re-calculation of the UBCS statistics defined by Dreszer *et al.* (2007). The estimation procedure of the expected number of the so-called clustered substitutions was modified through the introduction of the inclusion-exclusion principle. Our approach allows to calculate the exact value of UBCS statistic even for the complex structures of the intersecting clusters, which was unattainable with the original method. Consequently, we calculated the UBCS statistics for the Great Apes family and the updated estimation of the HSA2 fusion time. Furthermore, we discuss how the UBCS statistics can be used to derive evolutionary distances within the Great Apes family. Finally, we present an observation on the linearity of the number of biased clustered substitutions (BCS) occurrences with respect to time.

In the following section, we introduce the genomic datasets used in this study, i.e. the Great Apes, and modern humans. We then describe in detail the UBCS statistics and discuss its deficiencies and potential oversights. Next, we comment on the introduced changes in the UBCS statistics and their impact on the estimation of the ancestral fusion time. We point out other observations regarding the evolutionary events related to weak-to-strong mutations. Finally, we discuss the possible improvements that could be implemented into our analyses, especially when the missing fragments of the Great Apes chromosomes are available.

3.1. Materials and Methods

To better estimate the times of HSA2 fusion and split of modern human and Great Apes, we used the latest builds of these genomes. We present the derivation of the formulas used for the calculation of the UBCS statistics and emphasize the differences in calculations of the substitutions clusters as well as estimation method of the fusion time along with the determination of its confidence interval.

3.1.1. Genomic Data

All of the sequences and alignment files of the modern human and Great Apes genomes used in this study were downloaded from the UCSC Genome Browser (`https://hgdownload.soe.ucsc.edu/downloads.html`) (Kent *et al.*, 2002).

1. The analyses of substitutions between the modern human and Great Apes genomes were based on the reciprocal best alignments of:
   - the hg38 human genome assembly (December 2013)
   - the Clint_PTRv2/panTro6 assembly of the chimpanzee (*Pan troglodytes*) genome (panTro6, University of Washington, January 2018)).
   - the Mhudiblu_PPA_v0 assembly of the bonobo (*Pan paniscus*) genome (University of Washington, May 2020).
   - the GSMRT3/gorGor6 assembly of the gorilla (*Gorilla gorilla*) genome (gorGor6, University of Washington, August 2019).
   - the Susie_PABv2/ponAbe3 assembly of the orangutan (*Pongo pygmaeus abelii*) genome (ponAbe3, University of Washington, January 2018).
   - the Nleu_3.0/nomLeu3 assembly of the gibbon (*Nomascus leucogenys*) genome (nomLeu3, Gibbon Genome Sequencing Consortium, October 2012).
2. The substitutions between the modern human and Great Ape genomes were classified using the February 2019 (Mmul_10) assembly of the Rhesus (*Macaca mulatta*) genome (rheMac10, The Genome Institute at Washington University School of Medicine) as an outgroup. For this purpose, we used the chain file *hg38.rheMac10.rbest.chain* and the reference sequence of the Rhesus genome (rheMac10).

Data processing and analyses as well as statistical procedures were conducted using scripts written in the Python and R programming languages. The principal pipeline was implemented as a Snakemake (Köster and Rahmann, 2012) workflow to make it reproducible and scalable. All scripts and Snakemake workflow files are publicly available at GitHub page: `https://github.com/bposzewiecka/tytus`.

Identification of single-nucleotide differences between the modern human and Great Ape genomes

The analyses of the biased clustered substitutions (BCSs) require a distinction between the types of substitutions within the specific genomes.

First, single-nucleotide differences (SNDs) between the modern human and Great Apes genomes where identified using the reciprocal best alignments (the human genome was the target, and the Great Apes genomes were queries). The reciprocal best liftover chain file was used to map human genome regions to its homolog in the Rhesus (*Macaca mulatta*) genome.

Next, based on the processing procedures suggested by Dreszer *et al.* (2007), SNDs between the modern human and Great Apes' genomes were filtered. An SND was discarded if one of the conditions in the 11-base pair (bp) window with the SND in the middle was met: (i) a deletion or an insertion was present, (ii) more than 2 differences between the target and query were found, (iii) the target sequence could not be lifted-over to the Rhesus (*Macaca mulatta*).

Finally, each resulting SND was classified into one of the following three groups: (i) derived in target, (ii) derived in query, or (iii) inconclusive. If the human and Rhesus genome nucleotides were the same, the SND was classified as derived in query. Conversely, if the Great Apes and Rhesus genome nucleotides were the same, it was considered as derived in target. Other substitutions were classified as inconclusive. If the Rhesus base was A or T and derived base was C or G, the SND was considered as a *biased substitution*.

Having prepared the classification of SND between genomes, we proceed with their clustering and calculation of the statistics that summarizes the local enrichment in *biased substitutions*. Below, we refer to SND as a *substitution*.

Dreszer *et al.* (2007) defined the UBCS statistics as the difference between the observed and the expected number of BCSs in each window of 1 Mb (referred to as a region) on an entire chromosome (all windows are disjoint). For this purpose, a substitution is considered to be a *clustered substitution* (CS) if it belongs to a 300 bp window with at least four other substitutions. Next, a CS is considered a BCS if it belongs to a window with at least 80% of weak-to-strong substitutions (Fig. 3.1). In this setting, the null-hypothesis assumes no relationship between the bias towards weak-to-strong substitutions and the clustering of substitutions.

Nonetheless, Dreszer *et al.* (2007) presented the method of computing the expected

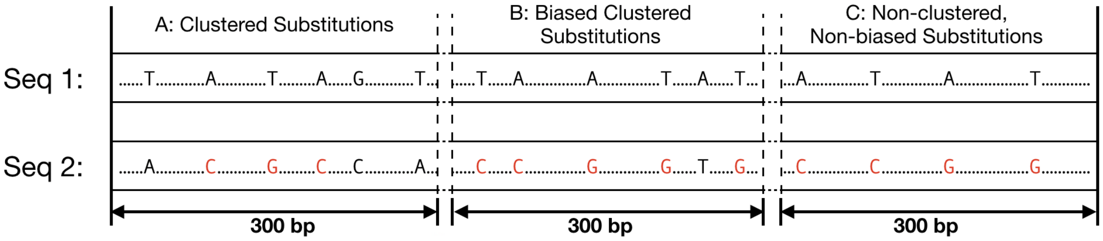

A: Clustered Substitutions
B: Biased Clustered Substitutions
C: Non-clustered, Non-biased Substitutions
Seq 1:
......T............A............T.........A.......G............T..
......T.......A...............A...............T.......A.......T...
...A...............T...............A..................T...........
Seq 2:
......A............C............G.........C.......C............A..
......C.......C...............G...............G.......T.......G...
...C...............C...............G..................G...........
300 bp
300 bp
300 bp

Figure 3.1: Examples of the substitutions classification for UBCS. The above figure depicts three 300 bp windows of two sequences, Seq 1 (reference) and Seq 2. Within each window substitutions that occurred on Seq 2 with respect to Seq 1 are denoted. The red color of the font is used for weak-to-strong substitutions. In the window A, all substitutions are considered as clustered substitutions (CSs), but not Biased Clustered Substitutions (BCSc), since only 50% of all substitutions are weak-to-strong. In the window B all substitutions are BCSs, because 5 out of 6 ($\geq$ 80%) substitutions are weak-to-strong. The remaining substitutions from the window C are neither clustered nor biased, because there are four substitution within this window.

.

number of BCSs only for a simplified case when one substitution can be included in at most 2 clusters. However, especially in the subtelomeric regions containing GC-rich isochores (Costantini *et al.*, 2006), the structure of the intersecting clusters can be more complex.

More precisely, Dreszer *et al.* (2007) relaxed the definition of CS by considering 300 bp windows that start at coordinates that are multiples of 150. In such a case, the computation of the expected number of BCSs simplifies, as at most 2 clusters sharing the same substitution have to be considered. Dreszer *et al.* (2007) provided an example of computation the probability that a substitution is BCS in one specific arrangement of substitutions in the overlapping bins. The method is based on the conditioning on the number of substitutions in the first cluster. Dreszer *et al.* (2007) do not provide any estimates of the complexity of their method.

Here, we have devised an efficient algorithm allowing for the computation of the UBCS statistics considering windows starting at coordinates that are divisors of the window's length. If a divisor is equal to 1, the algorithm during the computation of the probability that a given substitution is BC takes into account every possible window that the considered substitution is contained in. Such a procedure results in the precise calculation of the UBCS statistics by taking into account all possible window configurations of CSs. We also provide an estimation of the time and memory complexity of the described algorithm.

3.1.2. Efficient algorithm for the calculation of the expected number of BCSs

The expected number of BCS can be obtained by summing the probability of being BC for each substitution in the genomic region. Here we present an algorithm for computing the probability that a substitution is biased clustered (BC) given $\hat{p}$ and the arrangement of substitutions in all windows containing it. In the calculation of the expected number of

BCSs no association between bias and clustering is assumed. This quantity depends on the proportion of BS to all substitutions ($\hat{p}$) and the arrangement of substitutions in a genomic region.

Our algorithm compresses the genomic region containing each substitution into bins. Dynamic programming techniques allow for the computation of a probability that is tractable for the analyzed data in terms of time and computational memory consumption. To explain how the algorithm works, firstly we describe the procedure of compression of the genomic region containing the substitution in question into the vector of bins. Secondly, we present the derivations of formulas allowing for the application of the dynamic programming technique. Then, the pseudocode of the algorithm is shown. Finally, we explore the time and memory complexity of the algorithm.

The procedure of compression of the genomic region containing substitution into a vector of bins

Let us denote W as an event that a substitution at the coordinate j in the genome is BC and the respective probability as p'_j. To determine the value of p'_j all windows containing this substitution have to be considered as the potential biased clusters. Let m be a size of a window, and W_i an event that a window starting at a position $j - m + i$ is BC, where $i \in \{1, \dots, m\}$. The event W is a sum of the events that each window containing the coordinate j is BC, and can be expressed as:

$$p'_j = P(W) = P(W_1 \cup W_2 \cup \dots \cup W_m)$$

The number of the components of the sum needed to compute p'_j can be significantly reduced by unifying equal events and eliminating events with a zero probability. Therefore, from the windows that contain the same set of substitutions, only one representative can be left as a witness of being clustered. Windows containing less than 5 substitutions can be omitted, as their respective probabilities are zeros (because they are non-clustered).

Let us refer to the minimal set of windows that have to be considered in computing the probability of W after applying those rules as representative windows, and the number of such windows as n.

For the computation of a $P(W)$, a region covered by the representative windows can be compressed to a vector of size $2 \cdot n - 1$. Each element of such a vector represents a fragment of this region and stores the number of substitutions contained within it. Let us refer to each element of a such compressed representation as a *bin*. The coordinates of starts and ends of consecutive *bins* are determined by the ordered coordinates of starts and ends of all representative windows. The i-th cluster is defined as n *bins* starting at the i-th position

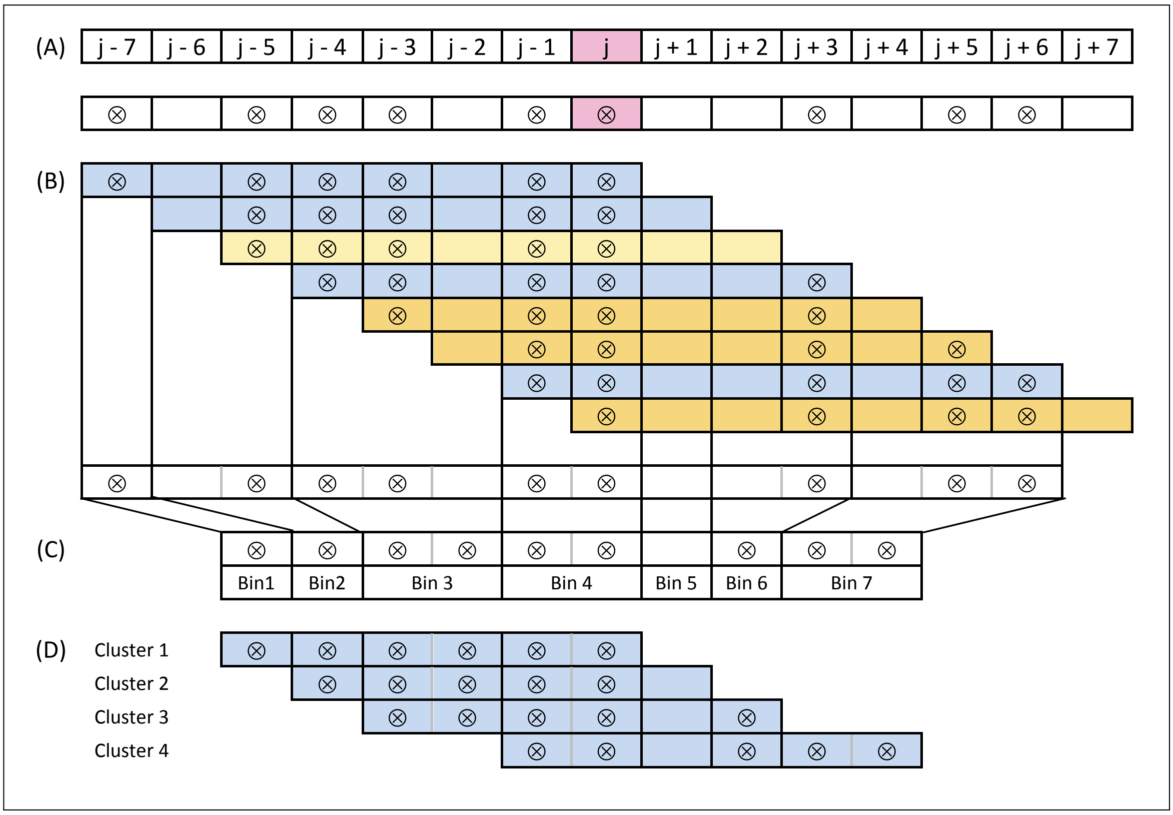


Figure 3.2: Illustration of the procedure of compression the genomic region containing a substitution into a vector of bins. The figure shows a compression procedure of a region of the genome containing a coordinate j and all windows of size $m = 8$ containing this coordinate. Substitutions are depicted as dots.

(A) Configuration of substitutions around the considered substitution at the j-th coordinate in the genome. All windows containing substitution from j-th coordinate (depicted in pink) cover a region from $j - 7$ to $j + 7$.

(B) Classification of all possible windows containing substitution at the j-th coordinate. All windows of size m containing the j-th coordinate are shown. Representative windows are depicted in blue. Window depicted in yellow is excluded because it contains the same set of substitutions as the preceding window. Windows starting depicted in orange are excluded as they contain less than 5 substitutions.

(C) Definition of bins based on all representative windows composition around the j-th coordinate. Starts and ends of representative windows (depicted by thick borders) mark the starts and ends of the bins. Note that substitution from the j-th coordinate is located in the middle bin.

(D) Resulting set of clusters corresponding to the representative windows. These clusters will be used for calculation of the probability that substitution from the j-th coordinate is the BCS (see Algorithm 3.).

and corresponds to the i-th representative window. This procedure of the compression windows into bins ensures that the substitution from the j-th coordinate is contained in the middle *bin*.

The upper bound for the cardinality of a minimal set of representative windows

Method of the construction of a minimal set of representative windows:

1. For each substitution from region starting at the $(j - m + 2)$-th and ending at the $(j-1)$-th coordinate, the first window that does not contain the previous substitution in the considered sequence is selected (in case there is no previous substitution we select the first window).

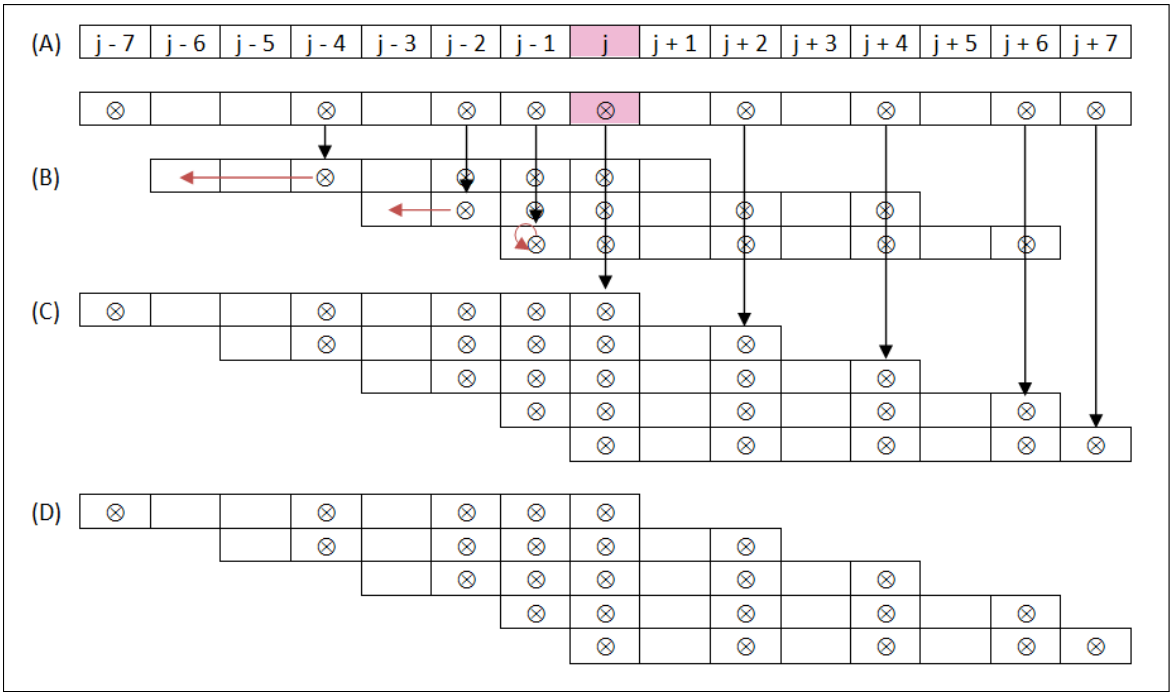


Figure 3.3: Illustration of the method of selecting representative windows. The figure shows a method of selecting representative windows from the region of the genome containing coordinate j and all windows of size $m = 8$ containing this coordinate. Substitutions are depicted as dots.

(A) Configuration of substitutions around the considered substitution at the j-th coordinate in genome. All windows containing substitution from j-th coordinate (depicted in pink) cover a region from $j - 7$ to $j + 7$.

(B) Candidates for representative windows corresponding to substitutions located in region from $(j - 6)$**-th to** $(j - 1)$**-th coordinate**. For each substitution from region starting at $(j - 6)$-th and ending at $(j - 1)$-th coordinate, the first window that does not contain the previous substitution is selected (indicated by the red arrows).

(C) Candidates for representative windows corresponding to substitutions located in region from j**-th to** $(j + 7)$**-th coordinate.** For each substitution from region starting at j-th and ending at $j + 7$ first window that contains this substitution is selected.

(D) Minimal set of representative windows. Windows with insufficient number of substitutions are removed from union of windows selected in (B) and (C).

2. For each substitution from region starting at the j-th and ending at the $j + m - 1$ the first window that contains this substitution is selected (the first window is always selected because it always contains substitution at the j-th coordinate).

3. From the set of windows selected in step 1 and 2 those with insufficient number of substitutions to satisfy BC conditions are removed.

To justify that events assigned to a set of windows selected in step 1 and 2 are equivalent to W, assume the opposite. Then, there must exist a window that is not equivalent to any of the selected windows and is not preceded by a substitution and does not end with a substitution. Let t be the starting coordinate of such window, and define l_j as the distance from coordinate j-th coordinate to the coordinate with the first substitution on the left of it. Window starting at $t - min(l_t, l_{t+m-1})$ coordinate is equivalent to the window starting at t, which contradicts with the initial assumption ($min(l_t, l_{t+m-1})$ is strictly greater then 0, since window starting at t-th coordinate is not preceded by a substitution and do not ends with the substitution).

In the method of the construction of the set, for each substitution at most one representative window is selected. Hence, the upper bound for the number of selected windows is equal to the number of substitutions in the region covered by all windows containing substitution at j-th coordinate.

Derivation of formulas used in the algorithm

The probability that s substitution form bin n is BC is equal to the probability of the sum of events that each cluster containing it is biased.

Let us denote:

- A as an event, that a substitution from n-th bin is BC (this substitution corresponds to the substitution located at j-th coordinate in the considered genomic region),
- A_k as an event, that the k-th cluster is biased,

$$p'_j = P(A) = P(A_1 \cup A_2 \cup \dots \cup A_n)$$

Notice that the event A corresponds to the previously considered event W and the selected representative windows W_i correspond to the clusters A_i and obviously $P(W) = P(A)$. Now, the formula for $P(A)$ can be written as a sum:

$$\begin{aligned} P(A) = P(A_1) + P(A_2 \cap \neg A_1) + P(A_3 \cap \neg A_1 \cap \neg A_2) + \dots \\ + \dots + P(A_n \cap \neg A_1 \cap \neg A_2 \cap \dots \cap \neg A_{n-1}) \end{aligned} \tag{3.1}$$

According to the law of total probability, for each k, every component of the above sum (3.1) of a form $P(A_k \cap \neg A_1 \cap \neg A_2 \cap \dots \cap \neg A_{k-1})$ can be expressed as:

$$\begin{aligned} \sum P(A_k \cap \neg A_1 \cap \neg \cap \dots \cap \neg A_{k-1} | X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}) \cdot \\ P(X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}) \end{aligned} \tag{3.2}$$

where X_k is a random variable specifying the number of the biased substitutions in the k-th bin, the summation is done for all $x_k \in \{0, b_k\}, \dots, x_{k+n-2} \in \{0, b_{k+n-2}\}$. Next, since both events A_k and $\neg A_1 \cap \neg A_2 \cap \dots \cap \neg A_{k-1}$ are conditionally independent given $X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}$, each component of the sum (3.2) is equal to the product of the following three terms:

$$P(A_k | X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}) \tag{3.3}$$

$$P(\neg A_1 \cap \neg A_2 \cap \dots \cap \neg A_{k-1} | X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}) \tag{3.4}$$

$$P(X_k = x_k, \dots, X_{k+n-2} = x_{k+n-2}) \tag{3.5}$$

The first term (3.3) specifies the probability that a cluster is BC, given the frequencies of the first $n-1$ bins. By the law of the total probability, it can be computed by conditioning on the frequency of the last bin in the cluster:

$$
\begin{aligned}
&P(A_k|X_k = x_k, \ldots, X_{k+n-2} = x_{k+n-2}) = \\
&= \sum_{x_{k+n-1} \in \{0, s_{k+n-1}\}} P(A_k|X_k = x_k, \ldots, X_{k+n-1} = x_{k+n-1}) \cdot \ P(X_{k+n-1} = x_{k+n-1})
\end{aligned}
$$

The value of the expression $P(A_k|X_k = x_k, \ldots, X_{k+n-1} = x_{k+n-1})$ indicates that the k-th cluster containing $\sum_{i=k}^{k+n-1} x_i$ biased substitutions is biased.

The second term (3.4) specifies the probability, that the first $k-1$ clusters are not biased, given the frequencies of the last $n-1$ bins of the $k-1$-th cluster. By conditioning on the frequencies of the $k-1$-th bin, this probability can be expressed using the low of total probability as:

$$
\begin{aligned}
&P(\neg A_1 \cap \cdots \cap \neg A_{k-1}|X_k = x_k, \ldots, X_{k+n-2} = x_{k+n-2}) = \\
&= \sum_{x_{k-1} \in \{0, b_{k-1}\}} P(\neg A_1 \cap \cdots \cap \neg A_{k-1}|X_{k-1} = x_{k-1}, \ldots, X_{k+n-2} = x_{k+n-2}) \cdot \\
&P(X_{k-1} = x_{k-1})
\end{aligned}
$$

Yet events, $\neg A_1 \cap \cdots \cap \neg A_{k-2}$ and $\neg A_{k-1}$ are conditionally independent given $X_{k-1} = x_{k-1}, \ldots, X_{k+n-2} = x_{k+n-2}$, thus:

$$
\begin{aligned}
&P(\neg A_1 \cap \cdots \cap \neg A_{k-1}|X_k = x_k, \ldots, X_{k+n-2} = x_{k+n-2}) = \\
&= \sum_{x_{k-1} \in \{0, b_{k-1}\}} P(\neg A_{k-1}|X_{k-1} = x_{k-1}, \ldots, X_{k+n-2} = x_{k+n-2}) \cdot \\
&P(\neg A_1 \cap \cdots \cap \neg A_{k-2}|X_{k-1} = x_{k-1}, \ldots, X_{k+n-2} = x_{k+n-2}) \ \cdot P(X_{k-1} = x_{k-1})
\end{aligned}
$$

The value of the probability $P(\neg A_{k-1}|X_{k-1} = x_{k-1}, X_k = x_k, \ldots, X_{k+n-2} = x_{k+n-2})$ indicates that the $k-1$-th cluster containing $\sum_{i=k-1}^{k+n-2} x_i$ biased substitutions is biased.

Finally, events $\neg A_1 \cap \neg A_2 \cap \cdots \cap \neg A_{k-2}|X_{k-1} = x_{k-1}, \ldots, X_{k+n-3} = x_{k+n-3}$ and $X_{k+n-2} = x_{k+n-2}$ are independent, thus $P(\neg A_1 \cap \neg A_2 \cap \cdots \cap \neg A_{k-1}|X_{k+1} = x_{k+1}, \ldots, X_{k+n-2} = x_{k+n-2})$ is equal to:

$$
P(\neg A_1 \cap \neg A_2 \cap \cdots \cap \neg A_{k-1}|X_{k-1} = x_{k-1}, \ldots, X_{k+n-3} = x_{k+n-3})
$$

Pseudocode of the algorithm

An algorithm for computing the probability that a substitution is BC is a straightforward application of the formulas derived above.

Computing all conditional probabilities given by the expression in Eq. (3.2) requires generating a Cartesian product representing all possible frequencies of biased substitutions in $n-1$ subsequent bins. A pseudocode of the recursive function GENERATE-BIN-FREQUENCIES is presented as Algorithm 1. This function returns a list of 2-tuples containing a list of frequencies together with their respective probabilities.

Algorithm 1:

```
GENERATE-BIN-FREQUENCIES($bin\_sizes, \hat{p}$)
    Data: A list of bin sizes and the probability that a substitution is biased
    Result: A list of 2-tuples containing all possible frequencies of biased
            substitutions bins and their respective probabilities
    if $length(bin\_sizes) = 0$ then
        return $list(tuple(list(), 1))$
    end
    result ← $list()$
    for $k \leftarrow 0$ to $bin\_sizes[1]$ do
        freqs_with_prob ←
          GENERATE-BIN-FREQUENCIES(SUBLIST($bin\_sizes, 2, n$), $\hat{p}$)
        for (freqs, prob) ∈ freqs_with_prob do
            new_freqs ← $list(k)$ + freqs
            new_prob ← prob · BINOM($sizes[1], k, \hat{p}$)
            result.append($tuple$(new_freqs, new_prob))
        end
    end
    return result
```

The function BINOM-FROM($bin\_size, start\_size, \hat{p}$) (Algorithm 2) returns the probability that a bin of size bin_size contains $start\_size$ or more biased substitutions, where $\hat{p}$ is the probability that substitution is biased.

The function GET-PROBABILITY-OF-BCS($bin\_sizes, \hat{p}$) (Algorithm 3) takes as the arguments a list of sizes of consecutive bins of all clusters that contain a given substitution and a probability that the substitution is biased. The function returns the probability that the substitution contained in the middle bin is BC.

In the lines 4-5 of the Algorithm 3, the first component of the sum (3.1) is computed. Next components are evaluated in $n-1$ iterations of the main loop in which the function GENERATE-BIN-FREQUENCIES is used for generation of all possible frequencies of the BSs in subsequent $n-1$ bins starting from the k-th bin.

Then, in the lines 16-18, the value of the conditional probability of the event that the

Algorithm 2:

```
BINOM-FROM($bin\_size, start\_freq, \hat{p}$)
  Data: Bin size (bin_size), start frequency (start_freq), and the probability of
        success ($\hat{p}$)
  Result: The procedure returns a sum of values of PMF of binomial distribution
          with parameters $bin\_size$ and $\hat{p}$ with the number of successes ranging
          from $start\_size$ to $bin\_size$
  result $\leftarrow 0$
  for $k \leftarrow start\_size$ to $bin\_size$ do
      result $\leftarrow$ result $+$ BINOM($bin\_size, k, \hat{p}$)
  end
  return result
```

cluster k is biased (term from Eq. (3.3)) is computed. In the lines 19-23, the value of the conditional probability that all previous clusters are not biased (term from Eq. (3.4)) is evaluated. For this purpose, the values from the dictionary $prev\_mem\_dict$ are used. The values in the dictionary mem_dict are updated for the use in the next iteration.

In the line 25, the result is updated by adding the product of the two probabilities (Eq. (3.3) and Eq. (3.4)) and the probability that $n-1$ bins contain the certain number of biased substitutions.

Analysis of the computational complexity of the algorithm

For the computation of the probability that a substitution is BC, the required memory is proportional to 2^c, where c is the maximum number of substitution in a clusters. The inner loop iterates over Cartesian product representing all possible frequencies of the biased substitutions in $n-1$ subsequent bins adding to the dictionary one value per iteration. Time complexity of the algorithm is proportional $n \cdot 2^c$ as main loop $n-1$ times iterate over Cartesian product defined above.

3.1.3. UBCS based evolutionary distance estimation

Finally, to determine whether and how the average proportion between the values of the introduced UBCS statistics for two genome sequences within both telomere regions correlates with the time of evolutionary speciation events, we derived the following UBCS proportion measure.

Specifically, let us assume that there are two genome sequences $\mathcal{G}_x$ and $\mathcal{G}_y$, N chromosomes and M windows of size 1 Mb on telomeric regions of each chromosome. We denote

$\mathcal{G}_{x^i_j}$ as j-th window on the i-th chromosome of the genome $\mathcal{G}_x$ and the value of its UBCS statistic as $\mathcal{U}(\mathcal{G}_{x^i_j})$, and $\overline{x}$ as the inverted sequence of x (i.e. the first window of $\overline{x}$ are the last 1 Mb of x). We calculate the average UBCS proportion between telomeres on p and q arms of i-th chromosomes of genomes $\mathcal{G}_x$ and $\mathcal{G}_y$ as:

$$\mathcal{T}_p(i) = \frac{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{x^i_j})}{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{y^i_j})} \qquad \mathcal{T}_q(i) = \frac{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{\overline{x}^i_j})}{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{\overline{y}^i_j})}$$

and the evolutionary distance based on the average UBCS proportion between genomes $\mathcal{G}_x$ and $\mathcal{G}_y$ as:

$$\mathcal{G}_x || \mathcal{G}_y = \text{median}(\{\mathcal{T}_p(i) : i \in \mathcal{CT}_p\} \cup \{\mathcal{T}_q(i) : i \in \mathcal{CT}_q\})$$

where $\mathcal{CT}_p$ and $\mathcal{CT}_q$ are sets of so called *control chromosomes* used to measure UBCS proportion between genomes on p and q arm respectively.

Such defined proportions allowed us to estimate the possible branching times in the evolutionary tree for each of the considered Great Apes genomes, that will be described in the next section.

For this purpose, we have computed the UBCS statistics using the human genome as a target and the Great Apes genomes as queries. Then, we have compared the distances between genome of the chimpanzee and all other Apes genomes by determining the value of an UBCS proportion $\mathcal{G}_x || \mathcal{G}_y$ defined above. We have used 10 windows of the size of 1 Mb, and the following sets of the control chromosomes $\mathcal{CT}_p = \{1, 4, 5, 6, 8, 10, 12, 16, 17, 19\}$ and $\mathcal{CT}_q = \{\text{all autosomes}\} \setminus \{15, 18, 19, 20\}$ for p and q arms, respectively. From the set of autosomal chromosomes, the short arms of the acrocentric chromosomes and the arms of chromosomes that were rearranged in human and Great Apes genomes were excluded. The confidence interval for the UBCS proportion was determined using the bootstrap method. The bootstrap sample was constructed by sampling with replacement of the 15 out of 28 telomeres, and 8 of 10 windows on the basis of which UBCS proportion is calculated. The sampling procedure was repeated 1000 times for each species, and confidence intervals were determined by eliminating 5% of the most extreme values. Speciation time was approximated by multiplication the UBCS proportion (quantifying the distance between the chimpanzee genome to genome of interest) by the estimated time of the human-chimpanzee split. We have fixed the human-chimpanzee speciation time at 6 Mya. The confidence intervals for the speciation events were obtained by rescaling the confidence intervals of the UBCS proportions in the same manner. For the purpose of more informative visualizations, in all of the figures regarding UBCS statistics, the `loess` regression function was used to smooth the curves.

3.1.4. UBCS based estimation of the fusion time

Our method of estimation of the HSA2 fusion time is based on two natural assumptions. First, analogously to Dreszer *et al.* (2007), we assumed a constant evolutionary force that has lead to the accumulation of BCS near telomeres in each species. Interestingly, our analysis of all Great Apes genomes revealed that the rate of this accumulation differs among species. The second assumption considers the time of human-chimpanzee split at approximately 6 Mya.

The method of the fusion time estimation proposed by Dreszer *et al.* (2007) used also a third assumption that the ratio of UBCS between the p and q arms of any chromosome is similar for human and chimp. This assumption is clearly violated in the data and therefore we have devised a different estimation procedure.

For the calculation of the fusion time, let us define R as a proportion of time of the last 6 million years that two chromosomes were not fused. Then, the fusion time can be estimated as $6\ \text{Mya} \cdot (1 - R)$. We can approximate R as a ratio of two quantities: the proportion of UBCS values in the region right next to the fusion site (homologous to the chimpanzee p arm on chromosome 2B) and the UBCS proportion $\mathcal{G}_x||\mathcal{G}_y$, comparing the statistic values computed using the substitution derived in human and in chimpanzee since the divergence from common ancestor. The former proportion, reflects the decline in the accumulation of BCS after the fusion event, the latter accounts for the differences in the rate of accumulation of BCS between human and chimp. For this estimation we have used the following sets of control chromosomes $\mathcal{CT}_p = \mathcal{CT}_q = \{1..12, 16, 17\}$. To increase the robustness of the procedure we repeated the calculation for the telomeric regions of different sizes (from 10 Mb to 25 Mb). The final evaluation of the fusion time used a median value of the proportions.

3.1.5. Robustness of the UBCS statistics

Comparison of the UBCS statistics for different number of overlaping windows

Figure 3.4 shows that the increasing number of windows considered in the calculation of the UBCS statistics, results in higher values of this statistic near telomeres. This shows that the extent of biased gene conversion phenomenon is captured more precisely when we consider more windows that substitution can be contained in. The Fig. 3.4 is accompanied with the first three rows of Table 3.1, were estimates for corresponding number of windows are presented.

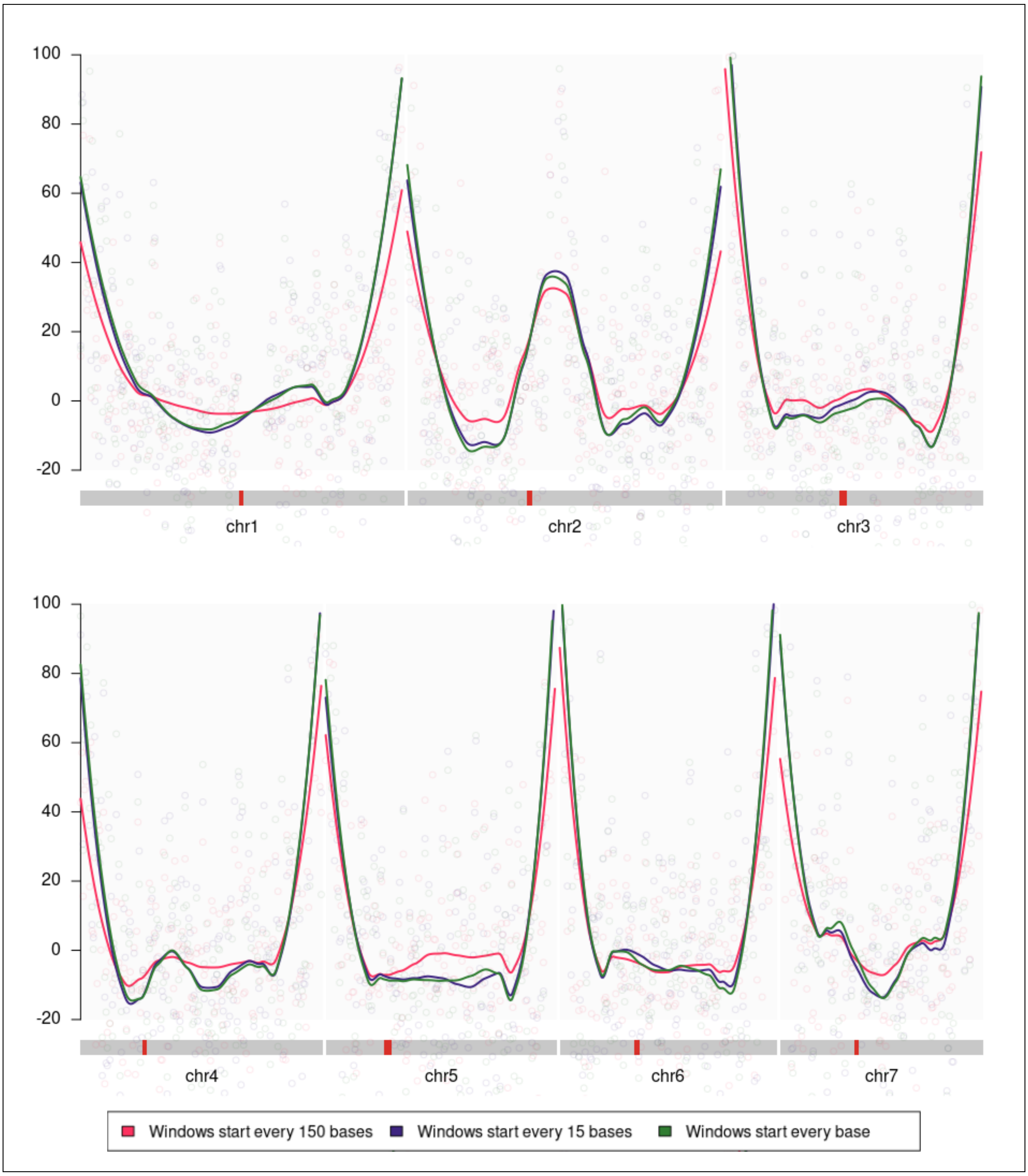


Figure 3.4: UBCS statistic for 7 longest autosomal chromosomes for different number of overlapping clusters. UBCS statistic is computed for windows of 300 bp that start every: (i) 150 bp - one substitution can be contained in 2 windows, (ii) 20 bp - one substitution can be contained in 15 windows, (iii) 300 bp - one substitution can be contained in 300 windows.

Comparison of the UBCS statistics for different definitions of BCSs

To assess the robustness of the algorithm with respect to its parameters we have evaluated the data for human and chimpanzee for different values of window sizes: 250 and 300; minimal percent of substitutions in window to consider a substitution *biased*: 75%, 80%, 83%; number of substitutions in the window to be considered *clustered*: 5 and 6. First of all the trends of the UBCS statistics are conserved for all telomere sites (significant increase of values) as well as around the fusion site of the Chr2. Additionally, we observe a low magnitude standard deviations from the values presented in the main article Fig. 3.5 shows UBCS values for selected set of parameters. The exact estimates with confidence intervals are also presented in the last four rows of the Table 3.1.

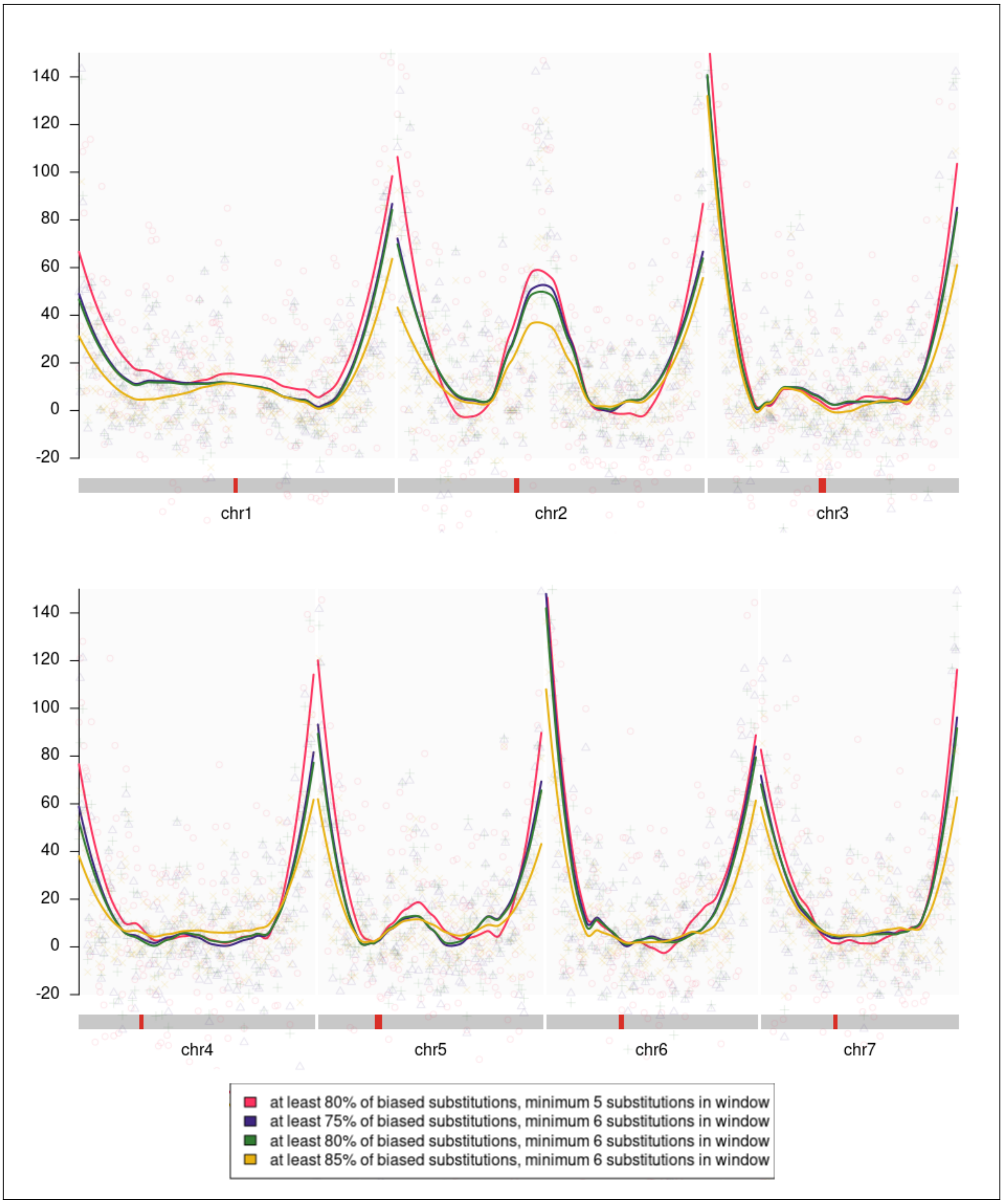


Figure 3.5: UBCS statistic for all 7 longest autosomal chromosomes for different definitions of BCSs. Substitution is defined as BCS if belongs to a window of 300 bp: (i) with at least 5 substitutions with at least 80% of weak-to-strong substitutions, (ii) with at least 6 substitutions with at least 75% of weak-to-strong substitutions, (iii) with at least 6 substitutions with at least 80% of weak-to-strong substitutions, (iv) with at least 6 substitutions with at least 85% of weak-to-strong substitutions.

3.2. Results

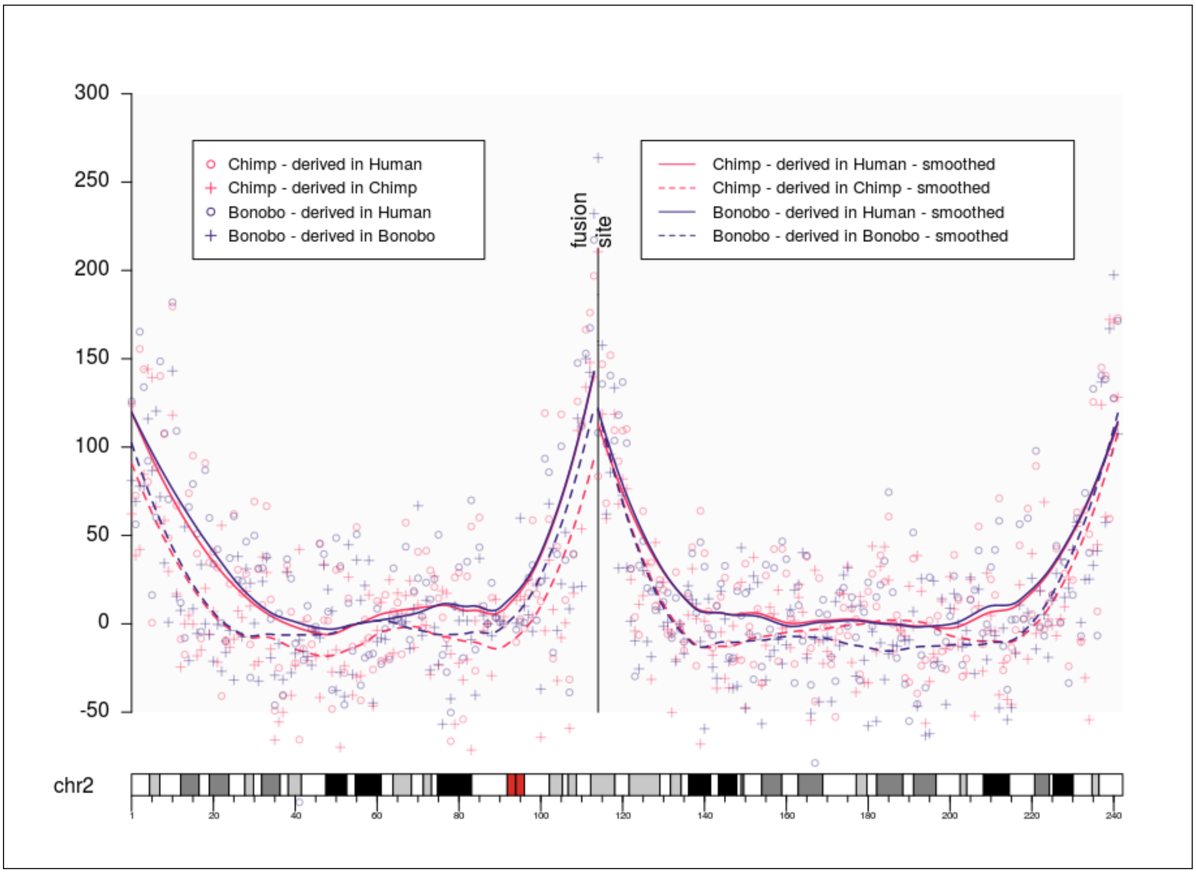


Figure 3.6: UBCS statistic for human chromosome 2 The figure above presents the values of the UBCS statistics along the whole chromosome 2. The vertical line denotes the ancestral fusion site point (chr2:113,500,000). One can observe how the lines corresponding to the same organism (solid vs dashed) differ from each other settling the difference between time estimation of the ancestral evolutionary split.

We present a revised estimation of the ancestral HSA2 fusion date based on the modified UBCS statistics. Furthermore, we present how the statistics corresponds to the evolutionary distances between human and Great Apes. Using the UBCS proportion between species, we have calculated the rates in which the BCS occurred in the telomeric regions. We have then used them to predict the timeline of the evolutionary events in the human lineage.

3.2.1. Revised HSA2 fusion date

First, after Dreszer *et al.* (2007), we have applied the UBCS statistics using the single nucleotide differences with a region of orthology in chimpanzee (*Pan troglodytes*). Additionally, we have added its evolutionary relative bonobo (*Pan paniscus*) to verify whether the UBCS statistics are consistent as might be expected in the context of evolutionary research (Hey, 2010).

In Fig. 3.6, we present the UBCS statistics values for both species that clearly indicate the HSA2 fusion site. Consequently, we have re-estimated the ancestral fusion date using the comparisons between the chimpanzee and modern human genomes to approximately 0.9 Mya with a 95% confidence interval of 0.4 - 1.5 Mya.

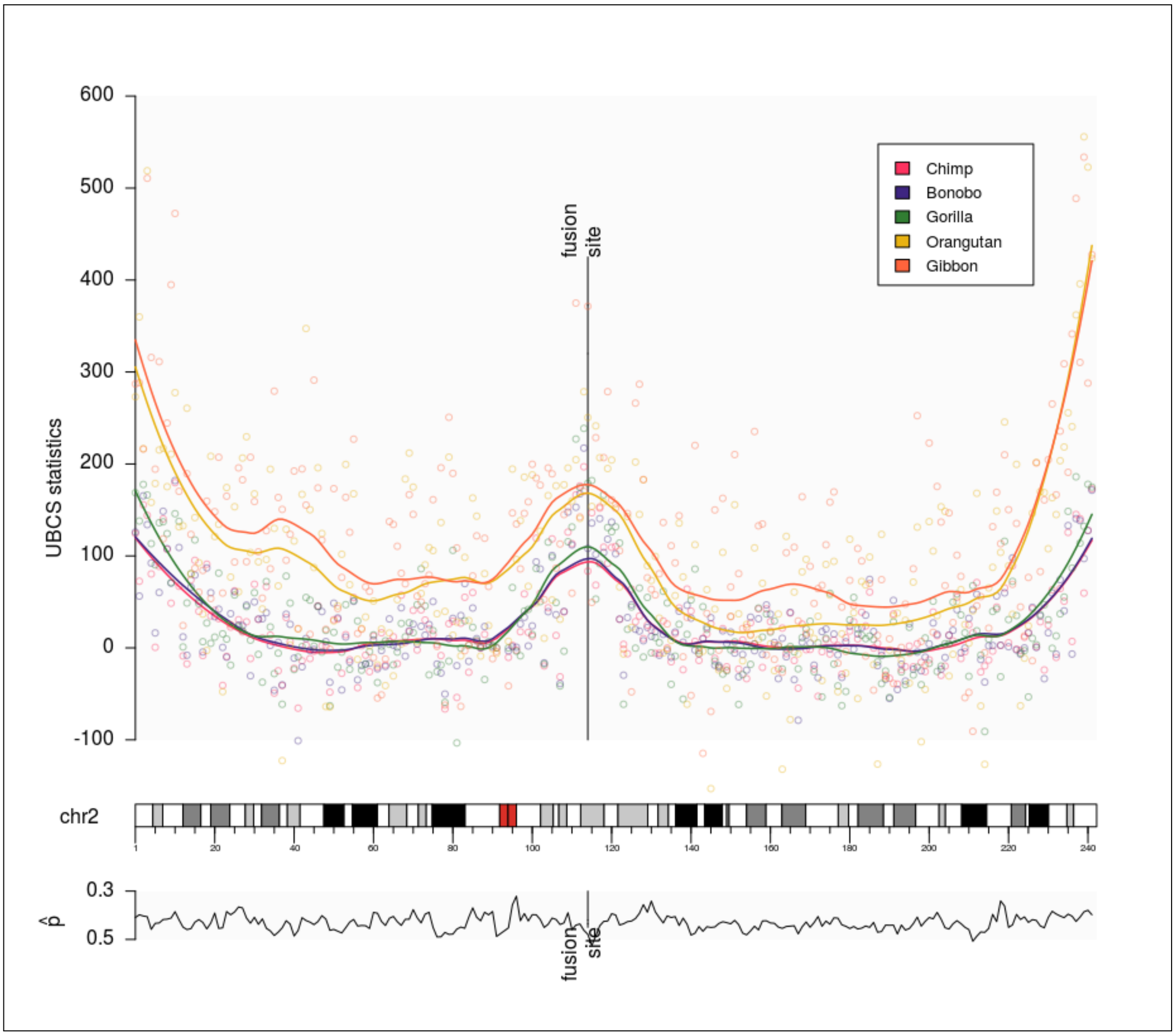


Figure 3.7: UBCS statistic for human chromosome 2 and Great Ape genomes (Top panel) UBCS statistic for substitutions derived in human are depicted as dots. Lines are the UBSC statistic values smoothed using loess regression. HSA2 shows the peak of the UBCS values near the ancestral fusion site. Atypical central peak occurs for the UBCS statistic computed using comparisons of the human to all Great Ape genomes. (Bottom panel) Values of $\hat{p}$ parameter (proportion of weak-to-strong in all substitutions) for every 1 Mb window of substitution derived in human chromosome 2.

Additionally, we have applied the same procedure of the fusion time estimation to the pair of the bonobo (*Pan paniscus*) and the modern human genomes. Since currently it is assumed that the present-day bonobo species have diverged from the common ancestor with modern human at the same time as chimpanzee (Hey, 2010), we expected that the estimation of the HSA2 fusion time will be similar to the one calculated based on the chimpanzee genome. Nonetheless, a time point was estimated as 0.67 Mya with 95% confidence interval 0-1.3 Mya. On one hand, this result contradicts the evolutionary reports. On the other hand, we observed a clear difference between the mutational dynamics of BCS on both sides of the fusion site. The proximal side maintains full compatibility between species, while on the distal side there is a double difference between species. In the next chapter, we discuss the possible reasons of these differences.

3.2.2. Coincidence of UBCS and evolutionary distances among Great Apes

Similarly as above, we have applied the UBCS statistics using single nucleotide differences within a region of homology to three more *hominidae* species: gorilla (*Gorilla Gorilla*), orangutan (*Pongo pygmaeus abelii*), and gibbon (*Nomascus leucogenys*). We show that for all five species, the UBCS statistics is monotonic as a measure of evolutionary distance (i.e. that species that are more evolutionary distant from human, have speciated prior to the others that have higher values of this statistics).

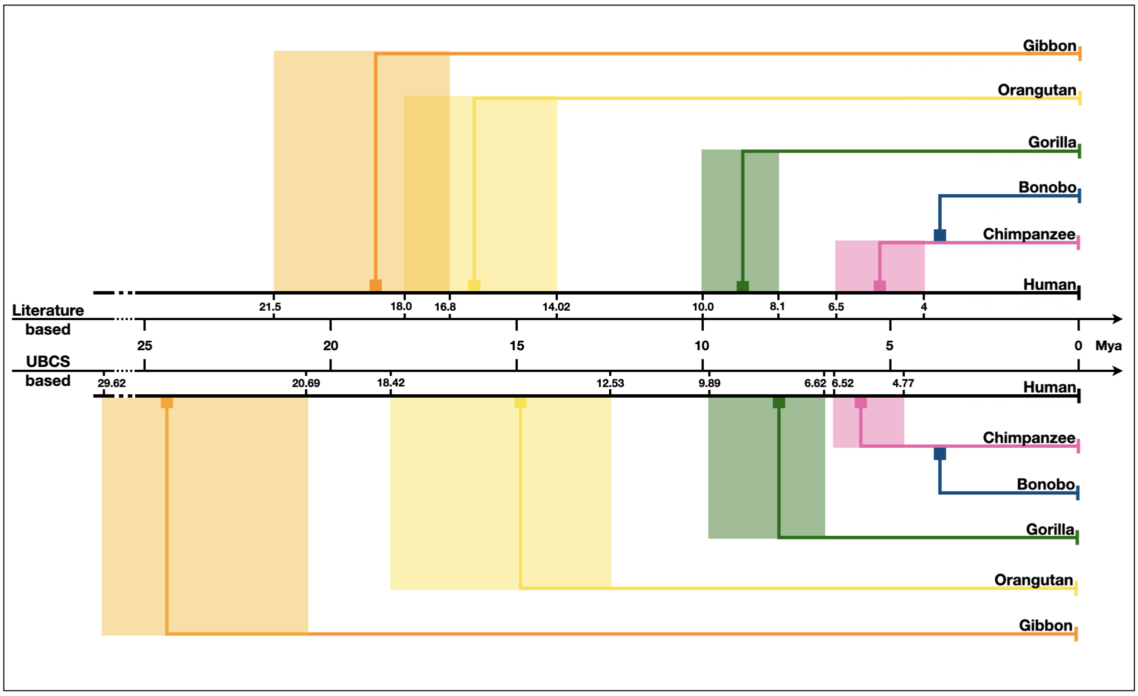


Figure 3.8: Evolutionary distances between Great Apes and Human. All recent reports about the possible speciation events times are shown. For each species, the minimal and the maximal dates are denoted on the horizontal time axis. Using the UBCS statistics proportions, we have estimated the time of the following divergence events from the human lineage for all species: Chimpanzee: 4.77-6.52 Mya, Bonobo: 4.35-5.85 Mya, Gorilla: 6.62-9.89 Mya, Orangutan: 12.53-18.42 Mya, Gibbon: 20.68-29.62 Mya. Please note, that each period calculated with the timeframe overlaps the with time frames taken from the literature.

Furthermore, based on the observation that the telomeric values of the UBCS statistics are consistently elevated for all autosomal chromosomes among all Great Apes (see Fig. 3.9), we have searched for the irregularity pattern. We have studied the relationship between the enrichment of BCS, thus values of the UBCS statistics, and evolutionary distances between organisms.

In the literature there are many reports aiming to estimate the speciation date of *hominidae* species from human (see Fig. 3.8). Starting chronologically, using Bayesian anal-

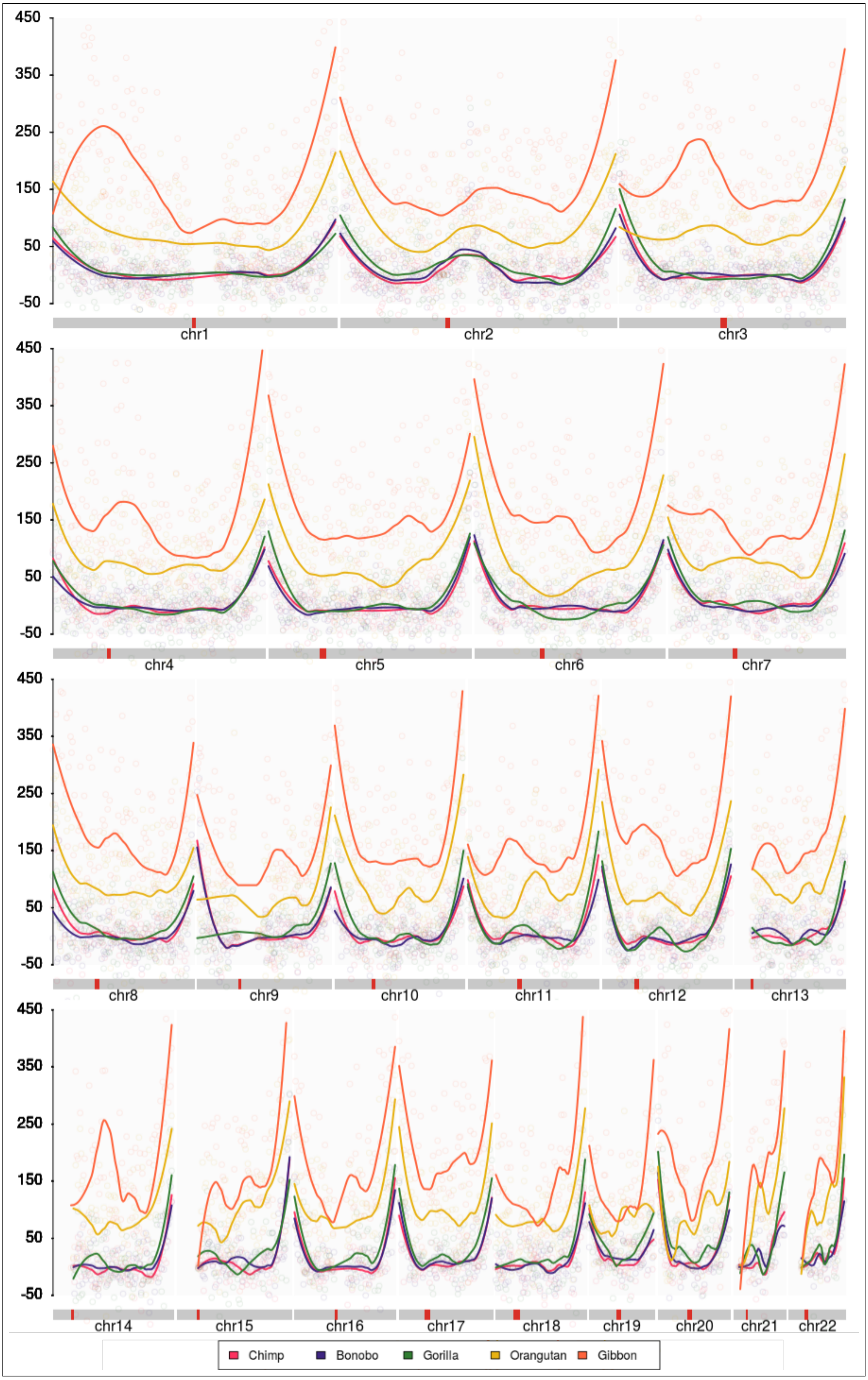


Figure 3.9: UBCS statistic for all autosomal chromosomes. The figure presents the value of the UBCS statistics over autosomal chromosomes for all five Great Apes studied in this chapter.

ysis with the relaxed clock model, the last common ancestor (LCA) of Gibbon *(Nomascus leucogenys)* and human was estimated by Chan *et al.* (2010) to have lived 19.25 Mya (95% confidence interval: 15.54–22.99 Mya). Using the relaxed clock model Chatterjee *et al.* (2009) estimated this event at 21.5 Mya (18.9-24.3). Carbone *et al.* (2014) suggested ĩ6.8 Mya (15.9-17.6) assuming a split time with macaque of 29 Mya and using the Bayesian coalescent-based methodology (Gronau *et al.*, 2011). Next, Orangutan *(Pongo Pygmaeus Abelli)* was estimated to speciate 18 Mya (Satta *et al.*, 2004) by applying the maximum likelihood (ML) method to intron sequences of 20 different loci. Later, a split time of 14.02 Mya (12.24–15.89) was suggested by Chan *et al.* (2010) using the same method as for gibbons (Chan *et al.*, 2010). Chatterjee *et al.* (2009) provided an estimation of 15.9 Mya (13.7-18.3). Speciation of Gorilla *(Gorilla Gorilla)* population by Chan *et al.* (2010) was estimated at 8.95 Mya (6.95–11.08). Raaum *et al.* (2005) suggested 8.1 Mya (7.1–9.0), whereas Scally *et al.* (2012) based on assembly and analysis of a genome sequence and fossil evidence places the specialization event at approximately 10 Mya. Further, based on coalescent hidden Markov model framework using in the context of incomplete lineage sorting, the existence of the LCA of chimpanzee and human was estimated at 6 Mya by Scally *et al.* (2012), 4 Mya by Hobolth *et al.* (2011) and 6.5–4.2 Mya by Stone *et al.* (2010) (see also references therein).

Based on the cited literature reports describing the estimations of the LCAs between various species and the over-representation of BCS near telomeric regions, we have found a specific relationship between UBCS statistics proportion and evolutionary distances for two given species. Using the method described in the subsection 2.5, we have calculated the speciation time for each pair of species based on their UBCS proportion $\mathcal{G}_x||\mathcal{G}_y$. For each species, for both minimal and maximal speciation time from the literature, we have estimated the average speciation value with respect to other species. As a result, we report the predictions of the speciation dates for all successive species. Chimpanzee and bonobo were are estimated to diverge very close to each other, between 4.7-6.6 Mya and 5.5-7.5 Mya, respectively. For gorilla, orangutan, and gibbon, the estimates are, respectively, 6.6-9.9 Mya, 12.5-18.4 Mya, 20.7-29.6 Mya. Overall, our results are consistent with the literature reports; however, the elder two species have a bit less robust estimation (see Fig. 3.8). In the Discussion section, we comment on the quality of this estimation as well as possible future improvements.

3.3. Discussion

Here, we provide a revised method for calculation of the UBCS statistics proposed by Dreszer *et al.* (2007). We have re-calculated the time of the HSA2 fusion event at approximately 0.9 Mya (0.4-1.5 Mya), using the same human and chimpanzee genomes comparison.

To verify our approach, we have used the bonobo genome as query because of their common evolutionary history (Hey, 2010). Interestingly, our results suggest that the fusion might have occurred more recently, approximately 0.6 Mya. We propose that this discrepancy may result from the quality of the bonobo genome assembly. Using the UCSC Browser (Kent *et al.*, 2002), we have observed that the genomic region distal to the HSA2 fusion site maps well to the near-telomeric region of chimpanzee chromosome 2B, and thus the corresponding UBCS statistics has high values (Fig. 3.6). Conversely, the genomic region proximal to the fusion site maps to the ambiguous region surrounded by closely located centromere and the large sequence gap. This observations may explain that the HSA2 fusion had a head-to-head type, but likely a big telomeric and sub-telomeric portion containing genes was lost (Stankiewicz, 2016).

Furthermore, we draw the reader's attention to the speciation estimation among the Great Apes. The short literature review described in the previous section presents how imprecise these estimations are. The differences in the calculated dates of the speciation events span from 2.5 Mya (chimpanzee) up 5 Mya (gibbon), demonstrating how challenging they are. We provide an evidence, that the UBCS statistic tracks a characteristic property of the human genomics, similar to the GC-content and consequently the BGC pattern (Romiguier and Roux, 2017; Meunier and Duret, 2004; Strathern *et al.*, 1995) and, can provide more accurate dating. It should be also noted that the evolutionary distance of *Hylobatidae* and *Ponginae* from modern *Homo sapiens* are substantial, that predictions based only on one type of data become rather blurred and imprecise. A remedy to that could be to use multi-layer models that would bring together various types of genomic and other -omic data (Marques-Bonet *et al.*, 2009b; Pinu *et al.*, 2019).

More recently, mapping the sequenced reads from modern humans and ancient Hominini (French, Han, Papuan, San, Yoruba, *Neandertal*, *Denisovan*) to the chimpanzee reference sequence (pantro2 version) facilitated more precise speciation events dating (Reich *et al.*, 2010). Quality scores given by Burrows-Wheeler Aligner (Li and Durbin, 2009) and ANFO (`https://bioinf.eva.mpg.de/anfo/`) software packages for mapping low-divergent sequences against a large reference genome that aim to reflect the confidence of its mapping to the chimpanzee genome have been used. Further adequate thresholds and restrictions to filter out the tentative nucleotides were applied. For the remaining data, the total number of transversion substitutions between all possible pairs of organism samples was counted. Finally, correction of the genetic divergence for sequencing error was estimated and revealed two principal observations: (i) the pairwise comparison of divergence results between 7 *Hominins* suggest that *Neandertal* and *Denisovan* are on average genetically related to each other more than either of them is related to modern humans; (ii) assuming human-chimpanzee genetic divergence at 6.5 Mya *Neandertal* and *Denisovan* divergence from a common ancestor was estimated to 644,000 years ago, while the divergence of both *Neandertals* and *Denisovans* to present-day Africans was estimated to 812,000 years ago.

These results are consistent with the reports by Green *et al.* (2010) who presented a draft sequence of the *Neandertal* genome. Using the numbers of transversions on the human lineage and the *Neandertal*-human ancestor to chimpanzee lineage the average divergence between DNA sequences in *Neandertals* and present-day humans, it was estimated as a percentage of the lineage from the modern human reference genome to the common ancestor of all considered organisms (i.e *Neandertals*, modern humans, and chimpanzees). The final estimate for the average divergence of *Neandertal* and modern human autosomal DNA sequences was estimated at 825,000 years ago, assuming the same human-chimpanzee split time.

3.4. Conclusions and Further Research

Herein, we aimed to aggregate the available genomic knowledge about the Great Apes species in order to provide more accurate estimation of the HSA2 chromosomal fusion time. We used an improvement of the approach described by Dreszer *et al.* (2007). We point out the drawbacks of their UBCS statistic and propose the improvements that made it more robust to parameter changes as well as taking into account the cardinality of the repetitive weak-to-strong substitutions within the analyzed scope. Finally, we provide the time estimations of the major speciation events that have occurred on the human lineage.

A possible extension of the presented work is to analyze the *Hominini* genomes. We intend to estimate the speciation events of *Denisovans* and *Neandertals* based on the UBCS statistics. Another interesting task would be to use more sophisticated way to estimate the evolutionary distances among Great Apes utilizing UBCS statistics with an incorporation a formal statistical model. The aim would be to make use of the theory of Hidden Markov Models (e.g. as presented in Hobolth *et al.* (2007, 2011)) or to formulate a Bayesian, coalescent-based model, e.g. as the one by Gronau *et al.* (2011).

```
GET-PROBABILITY-OF-BCS(bin_sizes, $\hat{p}$)
Data: A list of consecutive bin sizes of clusters containing substitution and the
      probability that substitution is biased
Result: The probability that the substitution from the middle bin is biased
        clustered
$n \leftarrow$ length(bin_sizes) $/\ 2 + 1$
result $\leftarrow 0$
first_cluster_size $\leftarrow$ SUM(SUBLIST(bin_sizes, $1, n$))
start_size $\leftarrow$ CEIL($0.8\cdot$ first_cluster_size)
result $\leftarrow$ BINOM-FROM(first_cluster_size, start_size, $\hat{p}$)
mem_dict $\leftarrow$ DICTIONARY(1)
                                        /* dictionary returning 1 by default */
for $k \leftarrow 2$ to $n$ do
    cond_sizes $\leftarrow$ SUBLIST(bin_sizes, $k, k + n - 2$)
    cluster_size $\leftarrow$ SUM(SUBLIST(bin_sizes, $k, k + n - 1$))
    prev_cluster_size $\leftarrow$ SUM(SUBLIST(bin_sizes, $k - 1, k + n - 2$))
    prev_mem_dict $\leftarrow$ mem_dict
    mem_dict $\leftarrow$ DICTIONARY(0)
                                        /* dictionary returning 0 by default */
    for (cond_freqs, cond_prob) $\in$ GENERATE-BIN-FREQUENCIES(cond_sizes, $\hat{p}$)
     do
        cond_freq_size $\leftarrow$ SUM(cond_freqs)
        n_a $\leftarrow 0$
        start_size $\leftarrow$ CEIL($0.8 \cdot$ cluster_size $-$ cond_freq_size)
        bin_size $\leftarrow$ bin_sizes [$k + n - 1$]
        a $\leftarrow$ BINOM-FROM(bin_size, start_size, $\hat{p}$)
        upper_bound $\leftarrow$ MIN(CEIL($0.8 \cdot$ prev_cluster_size $-$ 1), bin_sizes [$k - 1$])
        for $freq \leftarrow 0$ to upper_bound do
            mem_key $\leftarrow$ list(freq) + SUBLIST (cond_freqs, $1, n - 2$)
            n_a $\leftarrow$ n_a $+$ BINOM(bin_sizes [$k - 1$], freq, $\hat{p}$) $\cdot$ prev_mem_dict
              [mem_key ]
        end
        mem_dict [cond_freqs ] $\leftarrow$ n_a
        result $\leftarrow$ result $+$ a $\cdot$ n_a $\cdot$ cond_prob
    end
end
return result
```

Table 3.1: HSA2 formation estimates based on UBCS statistics. The table provides information on the estimated time of HSA2 formation with respect to different parameters used in the UBCS-based method. The Table can be cross-referenced with Fig. 3.4 and 3.5.

Number of bins within the windows size	Minimal number of weak to strong substitutions	Minimal percent of biased weak to strong substitutions (%)	Estimated speciation time (MYA)	95% Confidence Interval (MYA)
300	5	80%	0.8	0-2.02
15	5	80%	1.0	0-2.23
2	5	80%	1.3	0.31-2.35
300	6	83.33%	1.2	0-2.63
2	6	83.33%	1.5	0.33-2.66
300	6	75%	1.0	0-1.98
2	6	75%	0.9	0-2.35

4

An efficient algorithm for listing the Minimal Linear Eulerian Decompositions of the Karyotype Graphs

"In trying to count our many blessing
the difficulty is not to find things to count,
but to find time to enumerate them all."
— Aiden Wilson Tozer

As described in previous chapters, DNA rearrangements are important sources of structural changes that impact the evolution of species. Importantly, evolution is an ongoing process that shapes genomes of individuals and, consequently, affects their development and functioning. Novel types of sequencing technologies allow us to explore the enormous plasticity of genomes, particularly when it comes to *de novo* chromosomal rearrangements.

DNA rearrangements reshape a genome by breaking it in two or more segments and rejoining them in a different order, potentially resulting in the loss or gain of some fragments. The most common and well-studied genomic rearrangements are those with two breakpoints, such as deletions, duplications, inversions, and translocations. Others, involving more than two breakpoints, are termed complex chromosomal rearrangements (CCR). Over the last decade, new types of CCRs have been identified, characterized by a substantial number of breakpoints generated in a single mutational event. These phenomena are collectively known as chromoanagenesis in the scientific literature. They are categorized into the three groups based on the location of involved breakpoints and the mechanisms of their formation, namely chromothripsis, chromoplexy, and chromoanasynthesis. Chromothripsis was initially described in cancer (Stephens *et al.*, 2011) and subsequently observed in patients with congenital disorders (Kloosterman *et al.*, 2011; Collins *et al.*, 2017; Weckselblatt *et al.*, 2015), as well as unaffected individuals (De Pagter *et al.*, 2015). It represents a single catastrophic event involving one chromosome shuttering fragments of the genome, potentially resulting in the loss of DNA fragments. In this type of CCR, breakpoints cluster in one or a few chromosomal loci without any specific order and orientation. In the second type of massive CCR, called chromoplexy, breakpoint junctions can occur inter- and intra-chromosomally. This phenomenon was first observed in prostate cancer (Baca *et al.*, 2013) and recently reported in a healthy female carrying two *de novo* CCRs involving six chromosomes, with a total of 137 breakpoint junctions (Eisfeldt *et al.*, 2021). The hallmark of the third type of massive CCR, termed chromoanasynthesis, is a presence of copy number alterations along with copy-neutral fragments. This phenomenon in germline cells was first analyzed by Liu *et al.* (2011), and recently described in Nazaryan-Petersen *et al.* (2018) study. Both chromothripsis and chromoplexy are explained by the non-homologous end joining mechanism, which repairs double-strand breaks during chromatin disruption, whereas chromoanasynthesis is elucidated by replication-based mechanisms.

Chromothripsis and chromoplexy events do not result in the amplification of genetic material. Hence, if they affect only one of the two homologous chromosomes, the order and orientation of the rearranged fragments in the derivative chromosomes is determined unambiguously. In case of the chromoanasynthesis, however, the structure of the derivative chromosomes in most cases cannot be elucidated based solely on the rearrangement breakpoints. This is because the amplified fragments are longer than the available long-reads data and lack small polymorphisms, making it impossible to distinguish the amplified copies.

A model of CCR changing the amount of the genetic material, called Karyotype Graph (KG), has been proposed for cancer genomes (Aganezov *et al.*, 2019) and can be straightforwardly applied to the constitutional genome rearrangements. This model, however, does not provide information about the structure of the underlying derivative chromosomes, which hinders the comprehensive analysis of functional genomics changes caused by CCRs.

To address this issue, Aganezov *et al.* (2019) formulated the Minimal Eulerian Decomposition Problem (MEDP), which aims to find a collection of linear and/or circular derivative chromosomes with the minimal cardinality that can be inferred from KG . By implementing the concept of omnitigs (Tomescu and Medvedev, 2016) for KGs, the authors introduced the Consistent Contig Covering Problem (CCCP) to recover unambiguous contigs that can be inferred from KGs. Moreover, they have proposed a polynomial time algorithm for solving this problem and successfully applied it to the highly rearranged 17 prostate cancer genomes. However, in case of chromoanasynthesis, which is characterized by the errors in DNA replication, the resulting unambiguous contigs are often very short. This limitation highlights the necessity for an efficient algorithm that can enumerate all possible scenarios of such CCRs.

Listing of all object satisfying a specified property is one of the fundamental and extensively studied problems in combinatorics and graph theory. Examples of problems of this type include the enumeration of spanning trees (Shioura *et al.*, 1997), st-paths (Birmelé *et al.*, 2013), k disjoint st-paths (Grossi *et al.*, 2018) , cycles (Birmelé *et al.*, 2013), maximal cliques (Tomita *et al.*, 2006) and many others (Wasa, 2016). The output length of enumeration problems often grows exponentially with the size of the input, so the complexity of such problems is characterized in terms of both input and output size (output-sensitive). For enumeration algorithms there are several complexity classes, such as polynomial delay algorithms, where the time taken to generate two consecutive output solutions have to be polynomial in the input size.

In this chapter, we propose the enumeration algorithm for efficient listing all Minimal Linear Eulerian Decompositions of KG with an time delay of $O(log(n)^2 \cdot l)$, where n is a number of vertices in KG and l is a length of the decomposition. To this end, we traverse recursion tree in a way that avoids dead ends. This is accomplished by incorporating the concept of a certificate, which is a data structure that ensures the existence of at least one solution in the currently processed node of the recursion tree. Finally, we apply our algorithm to enumerate all possible rearrangement scenarios in case of a one proband with congenital chromoanasynthesis from the Nazaryan-Petersen *et al.* (2018) study.

4.1. Methods

4.1.1. Preliminaries

Given an undirected multi-graph $G = (V, E)$, we define the number of vertices in G as n and the number of edges as m. Each vertex is uniquely represented by a number from the set $\{1..n\}$ and referred to as a *vertex number*. The multiplicity is a function $\mu : E \rightarrow \mathbb{N}_{\geq 0}$ encoding edges copy number. The sum of all multiplicies of a graph, $\sum_{e \in E} \mu(e)$, is referred to as l. A non-trivial graph is a graph where $\sum_{e \in E} \mu(e) \neq 0$.

A *trail* is a sequence of adjacent vertices and edges, where the number of the occurrences of edges is less than or equal to their multiplicities. For $s, t \in V$, an st-trail, denoted as $s \rightsquigarrow t$ is a trail where the first vertex is s and the last vertex is t. Let τ be a $t \rightsquigarrow u$ trail, we denote $G \setminus \tau$ as a graph G with multiplicities decremented by the number of occurrences of the corresponding edges in the trail τ. For an edge $\{u, v\}$, we denote $t \rightsquigarrow u \cdot (u, v)$ as the extension of the trail τ with the new edge $\{u, v\}$.

Each trail can be represented as a sequence of ordered pairs of vertices, which we refer to as an *edge representation of a trail.* Let us define a *canonical trail* as a trail whose *edge representation* is lexicographically less than or equal to the *edge representation* of the reversed trail.

4.1.2. Model of complex chromosomal rearrangement

Here, we present a model of the CCRs called Karyotype Graph. This model was introduced by Aganezov *et al.* (2019) and therefore in our definitions we follow the nomenclature used therein.

Let the *reference genome* be a set of *reference chromosomes* described by their names and lengths. An *extremity* is defined by a *reference chromosome*, coordinate, and type (tail or head). A *reference chromosome fragment* (RCF) is a pair of extremities from the same chromosome, with one being a tail extremity, T, and the other being a head extremity, H ($f = [f^T, f^H]$). A RCF represents a continuous part of the chromosome constrained by the two extremities. The coordinate of a *tail extremity* of an RCF is less than the coordinate of its *head extremity.* Each reference chromosome can be partitioned into a set of non-overlapping RCFs and each element of this partition is called a *segment.* An unordered pair of *head* and *tail extremities* from the same *segment* is referred to as a *segment edge.* An unordered pair of two *extremities* that is not a *segment edge* is called an *adjacency.* An *adjacency* formed by *head* and *tail extremity* from two consecutive *segments* in the *reference chromosome* is referred as *reference adjacency.* An *adjacency* that is not a reference adjacency forms a *breakpoint adjacency.*

Definition 1 (Karyotype graph (KG))**.** *A Karyotype Graph is a undirected multi-graph build upon the partition of the reference chromosomes into segments. A set of extremities establishes the vertices of the graph. The edges in this graph can be classified into two types: segment edges (encoding segments) and adjacency edges (encoding transitions between segments).*

Copy number excess $x(v)$ *on vertex* v is defined as the difference between the multiplicity of a segment edge containing v and the sum of the multiplicities of adjacencies incident to v, doubled in case of adjacency (v, v). A vertex v with a copy number excess greater then 0 is referred as a *telomere.*

Derivative chromosome is defined as a sequence of directed RCFs, where RCF in the forward direction is described by a pair $[f^T, f^H]$, and in the reverse direction as a pair $[f^H, f^T]$.

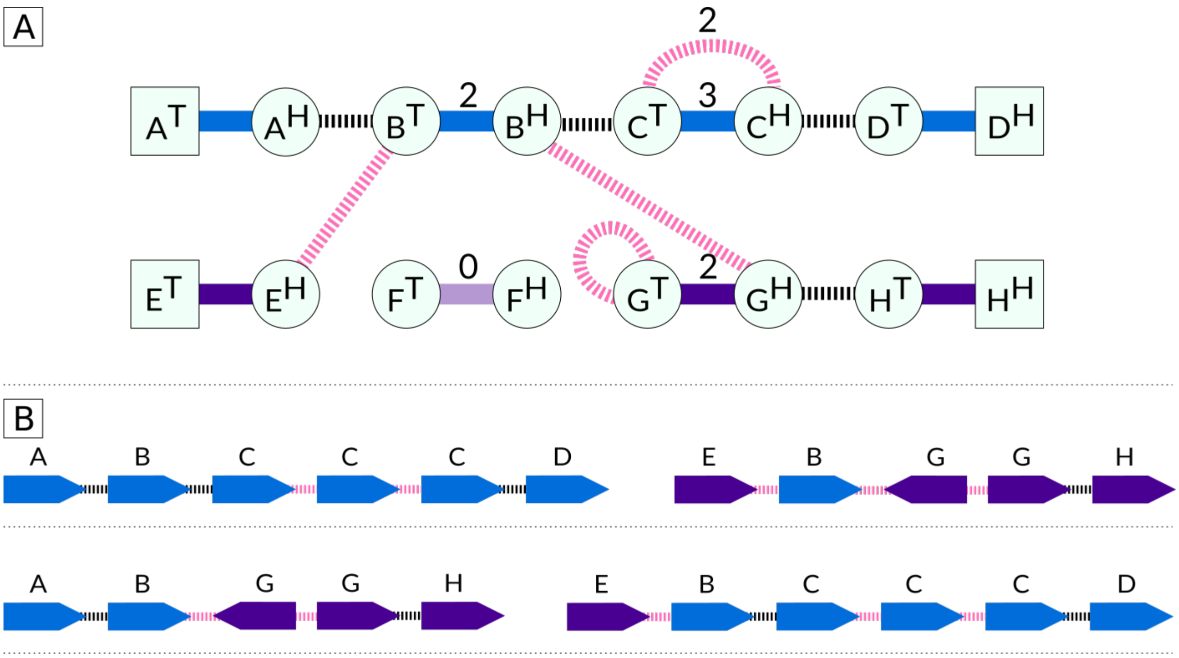


Figure 4.1: Linearly Decomposable Karyotype Graph and corresponding rearranged genomes. (A) The Linearly Decomposable Karyotype Graph depicts two reference chromosomes. The first reference chromosome consists of segments A, B, C, and D, while the second one of segments E, F, G, and H. Segment edges are represented as solid lines, with color corresponding to the reference chromosomes. Reference adjacencies are shown as black dashed lines, and breakpoint adjacencies are shown as pink dashed lines. Edges with copy numbers different from one are labeled accordingly. The deleted segment (F, with copy number 0) is depicted in faded color. Extremities are shown in light green boxes, with telomeres depicted as squares and other extremities as circles. Tail and head extremities are marked with superscripted T and H letters, respectively. **(B)** Two rearranged genomes are shown, each consisting of two derivative chromosomes corresponding to the Linearly Decomposable Karyotype Graph shown in panel (A). Adapted from Aganezov *et al.* (2019).

In such a sequence, two consecutive RCFs with the same orientation cannot originate from subsequent fragments in the reference genome (the sequence of RCFs describing the rearrangement should be as short as possible). *Rearranged genome* is defined as a sequence of *derivative chromosomes*. *Partition into segments inferred from a rearranged genome* is defined naturally.

The KG inferred from rearranged genome is defined by the set of extremities obtained from partition into segments. As defined above, the edges of a KG are of two types: adjacency edges and segment edges. Using partition into segments derived from rearranged genome, each RCF of the derivative chromosome is converted into a sequence of consecutive segments. Each segment in this sequence establishes segment edge, denoted as $[s_i^T, s_i^H]$, while the neighboring segment extremities reference adjacency edge $[s_i^H, s_{i+1}^T]$. Each pair of the neighboring extremities from an RCF forms one breakpoint adjacency edge. Therefore, the graph defined in such a way is a multigraph.

Every *derivative chromosome* is represented as a path consisting of alternating *segment* and *adjacency edges* starting and ending with a *segment edge*. Each *breakpoint adjacency* edge demarcates two RCF fragments represented by a sequence of alternating *segment edges* and *reference adjacency edges* starting and ending with a *segment edge*.

A *Linearly Decomposable Karyotype Graph (LDKG)* is a KG that can be inferred from

some *rearranged genome.* Every *rearranged genome* is represented by one LDKG, while a LDKG can represent several *rearranged genomes.*
An example of LDKG and corresponding linear rearranged genomes is presented in Fig. 4.1.

Definition 2 (Eulerian Decomposition of KG)**.** *An Eulerian Decomposition of KG is a sequence of alternating segment-adjacency linear trails and cycles in which every edge e is used exactly $\mu(e)$ times.*

Lemma 1. *KG has an Eulerian Decomposition if and only if for all $v \in V$, it holds that $x(v) \geq 0$ (Aganezov and Raphael, 2020; Aganezov et al., 2019; Oesper et al., 2012; Pevzner, 1995).*

Lemma 2. *Number of linear trails in any Eulerian Decomposition of KG is equal to $\sum_{v \in E} \dfrac{x(v)}{2}$ (Aganezov et al., 2019).*

Let us refer to the number of linear trails in Eulerian Decomposition of KG as k.

Lemma 3. *KG that has Eulerian decomposition is Linearly Decomposable if and only if it does not contain connected components without telomeres.*

Proof. If KG is Linearly Decomposable it cannot contain connected components without telomeres, because each Eulerian Decompositions of such component, by lemma 2, does not contain linear trails. If KG with Eulerian Decomposition does not have connected components without telomeres, any of its Eulerian Decomposition can be transformed to Linear Decomposition by merging iteratively cycles with linear trails using shared segment edges until there are no cycles.

Definition 3 (Minimal Eulerian Decomposition of LDKG)**.** *The Minimal Eulerian Decomposition of LDKG refers to the decomposition with the minimal cardinality.*

Throughout the article, we consider the *Minimal Eulerian Decomposition* of LDKG. According to the definition of LDKG this decomposition consists only of linear *derivative chromosomes.* Although, it is worth noting that there may exist a *Minimal Eulerian Decomposition* of certain KGs containing cycles representing circular chromosomes.

Definition 4 (Minimal Ordered Eulerian Decomposition of LDKG)**.** *The Minimal Ordered Eulerian Decomposition (MOED) of LDKG is a list of linear derivative chromosomes forming the Minimal Eulerian Decomposition of LDKG in their canonical representation and sorted in lexicographic order.*

All definitions, apart from LDKG and MOED, was taken from Aganezov *et al.* (2019).

4.1.3. Listing all distinct Minimal Ordered Eulerian Decompositions of LDKG

Definition 5 (Augmented Linearly Decomposable Karypotype Graph)**.** *Augmented Linearly Decomposable Karyotype Graph (ALDKG) is built upon the Linearly Decomposable Karyotype Graph by adding to it supplemental vertices s^T and s^H, corresponding supplemental segmental edge $\{s^T, s^H\}$ with the multiplicity $2 \cdot k$, and supplemental adjacency edge $\{s^T, s^T\}$ with the multiplicity $k-1$. Additionally, for every telomere t, an adjacency edge $\{t, s^H\}$ in LDKG is added, with the multiplicity equal to the copy number excess $x(t)$. The vertex number assigned to s^H is 0 and to s^T is $m+1$.*

The method of constructing ALDKG guarantees that it is connected and contains only one telomere s^T with *copy number excess* $x(s^T)$ equal to 2. Therefore, using lemma 2 and the fact that LDKG is connected, the MOED of ALDKG consists of one linear trail.

Let us redefine MOED for ALDKG by introducing additional constraint requiring that the subsequence of an edge representation of a trail, composed of all consecutive ordered edges starting with s^H is sorted using *vertex number* of the second element of the ordered edge. With such redefined notion of MOED for ALDKG, the enumeration of MOEDs in LDKG is in one-to-one correspondence betwwen the enumeration of MOEDs of ALDKG. The MOED of LDKG is obtained from MOED of ALDKG by removing from the only trail all edges incident to s^T and s^H.

Construction of ALDKG is similar to the construction of modified graph used in the proof of Theorem 1 in Aganezov *et al.* (2019). The approach to the the enumeration of MOEDs is similar to the procedure used in Grossi *et al.* (2018), although application in this context it is completely original.

4.1.4. Recurrent approach

Generating all possible MOEDs of ALDKG begins with exploring trails starting with supplemental segmental edge (s^T, s^H) and it can be viewed as a recursive scheme, where currently explored beginning of a trail $s^T \rightsquigarrow u$ is already fixed (initially $u = s^H$, where s^H is supplementary head vertex). It proceeds recursively by extending the trail $s^T \rightsquigarrow u$ at each step by adding the two consecutive edges (adjacency and segmental) to the trail. For this purpose, each *good neighbor* v of u is explored. A *good neighbor* of vertex u different from s^T is a vertex v such that $\{u, v\}$ is an adjacency edge and the reduced graph $ALDKG' \equiv ALDKG \setminus (s^T \rightsquigarrow u) \cdot (u, v)$ does not contain more than one non-trivial connected components. A vertex v is a *good neighbor* of s^T, if it is incident to s^T using adjacency edge and has the lowest *vertex number*. For each *good neighbor* v, the recursion proceeds with the extended trail $s^T \rightsquigarrow w \equiv (s^T \rightsquigarrow u) \cdot (u, v) \cdot (v, w)$, where $\{v, w\}$ is a segmental edge and u is set to w. The recursive procedure terminates when all edges are used up.

The recursion tree that is traversed while generating all possible MOEDs of ALDKG is presented in Fig. 4.2. Throughout the rest of the article we will use term nodes when referring to the recursion tree, and vertices when referring to the input ALDKG.

4.1.5. Recursion certificates

To make recursion efficient, two auxiliary data structures are introduced: the *witness certificate* and the *connectivity certificate.* The former is used to derive some properties of MOEDs, which helps in reduction of the time complexity of the algorithm, while the latter enables efficient queries preventing from the creation of dead ends in the recursion tree.

Witness certificate

Definition 6 (Witness certificate)**.** *For a given $LDKG$ the witness certificate is defined as any Minimal Ordered Eulerian Decomposition of LDKG. The trials in the witness certificate are referred to using their edge representation.*

The properties of the *witness certificate* are utilized to reduce the number of queries to the *connectivity certificate.* This is achieved by ensuring that the decrementing by one the multiplicity of only one adjacency edge incident to the processed vertex can result in dead end in the recursion tree.

Let us denote the i-th trail in *witness certificate* as T_i and the first edge of T_1 as (t, u). A positive answer to the question of whether some vertex v is a good neighbor of u can be given in the following cases:

1. pair (u, v) is the second element of the trail T_1.
2. pair (u, v) or (v, u) is an element of a trail T_i different from the T_1 (Fig. 4.3a, Fig. 4.3b).
3. pair (v, u) is an element of the trail T_1 (Fig. 4.3c).
4. pair (u, v) is an element of the trail T_1 and there exists another trail T_i with an edge containing vertex u (Fig. 4.3d, Fig. 4.3e).
5. pair (u, v) is an element of the trail T_1 and there exists another pair of vertices in T_1 containing u that is located further in a T_1 (Fig. 4.3f, Fig. 4.3g).

By applying appropriate transformations to the trails from the *witness certificate* we can obtain another Minimal Eulerian Decomposition of LDKG beginning with the segmental edge (t, u), followed by the adjacency edge (u, v) and containing the same multiset of edges.

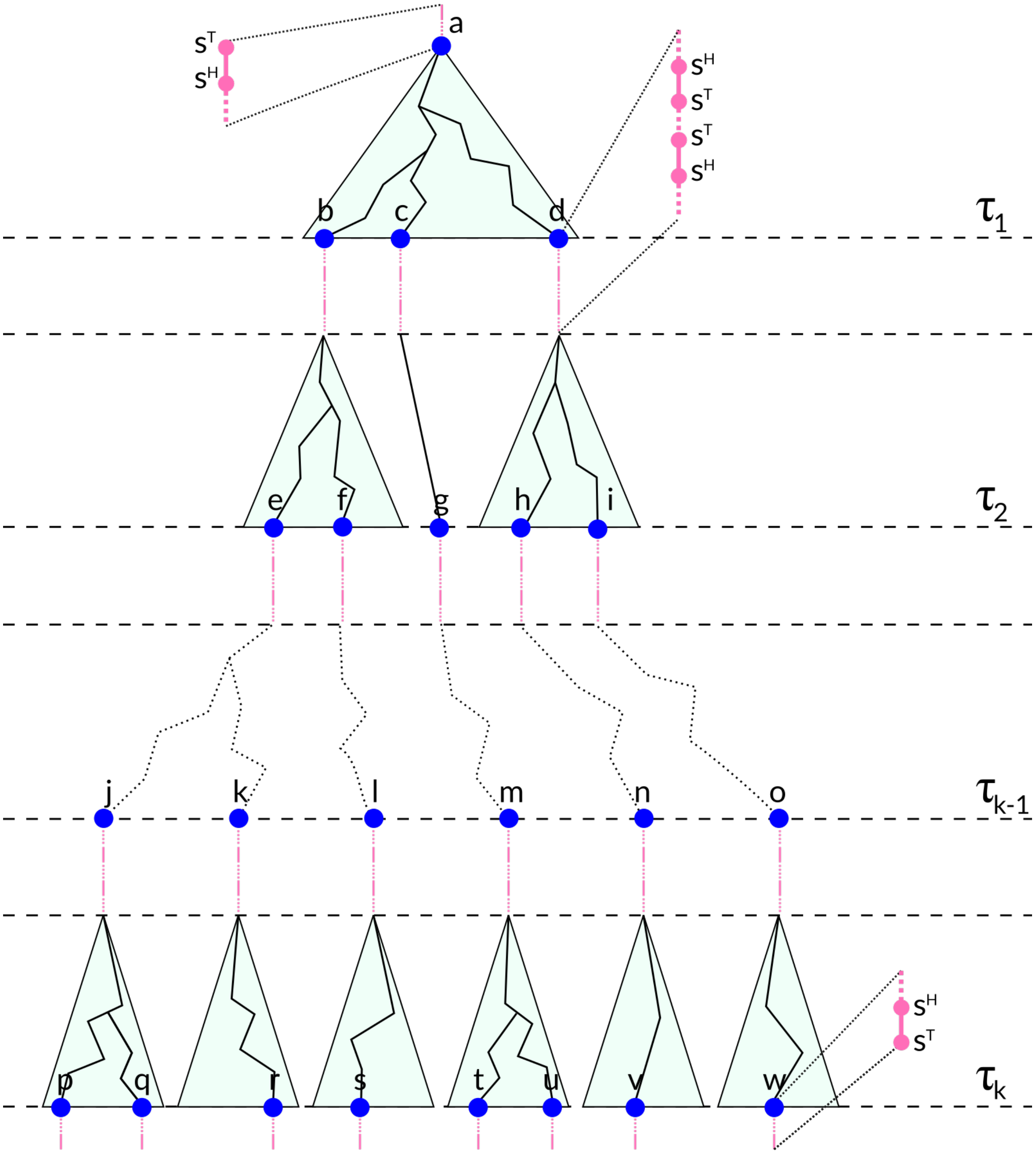


Figure 4.2: Recursion tree for enumerating Minimal Ordered Eulerian Decompositions of Augmented Linearly Decomposable Karyotype Graph. τ_i denotes the i-th trail in the MOED of LDKG corresponding to ALDKG. The recursion begins by extending a trail formed by a single supplemental segmental ordered edge (s^T, s^H) with a vertex that is a telomere with the lowest *vertex number* in the corresponding LDKG. The top tree is built of paths that represent all feasible choices of τ_1 and its leaves on the first dashed level correspond bijectively to trails τ_1. All leaves of the top tree are the roots of the recursion tree for unary nodes that generate paths corresponding to visiting vertices s^H and s^T and continuing the extension of a trail with a vertex that is a telomere with the lowest *vertex number* in corresponding LDKG that has not have been used. Traversing the unary node chain produces a fragment of a trail equal to $(t_1, s^H), (s^H, s^T), (s^T, s^T), (s^T, s^H), (s^H, t_2)$, there t_1 and t_2 are telomeres in the corresponding LDKG. Leaves in the third dashed levels correspond bijectively to trails τ_2, and so on. Nodes b, c, d correspond to possible choices of τ_1, whereas nodes e and f correspond to possible choices of τ_2 where τ_1 is represented by a path ending with node b (respectively g, when τ_1 is represented by c, next h and i, when τ_1 is represented by path ending with node d). Nodes from the last dashed line (from p to w) correspond bijectively to a sequence of trails $\tau_1, ..., \tau_k$ that form MOED of LDKG. The recursion ends by extending the trail with an adjacency edge incident to s^H and the supplemental segmental ordered edge (s^H, s^T).

Therefore, the only case when *witness certificate* cannot be used to confirm that the vertex v is a *good neighbor* of vertex u using simple transformations is when the pair (u, v) represents last occurrence of u in a trail T_1 and all of other pairs containing u belong to the trail T_1 but do not include the vertex v. Let us refer to such a vertex as *uncertain vertex.*

Connectivity certificate

Definition 7 (Connectivity certificate)**.** *A connectivity certificate represents a dynamic multi-graph initially equal to ALDKG. . It allows for the decrementing and incrementing the multiplicity of edges, as well as answering for the queries regarding the presence of more than one non-trivial connected component.*

The *connectivity certificate* is used to determine whether decrementing the multiplicity of a given edge from a multi-graph built upon ALDKG results in the formation of two non-trivial connected components. This data structure can be used to answer such questions by querying whether, after decrementing the multiplicity of the edge in question, the vertices forming the edge are not isolated and are connected to the supplemental vertex s^H through a trail with edges of multiplicity greater than 0.

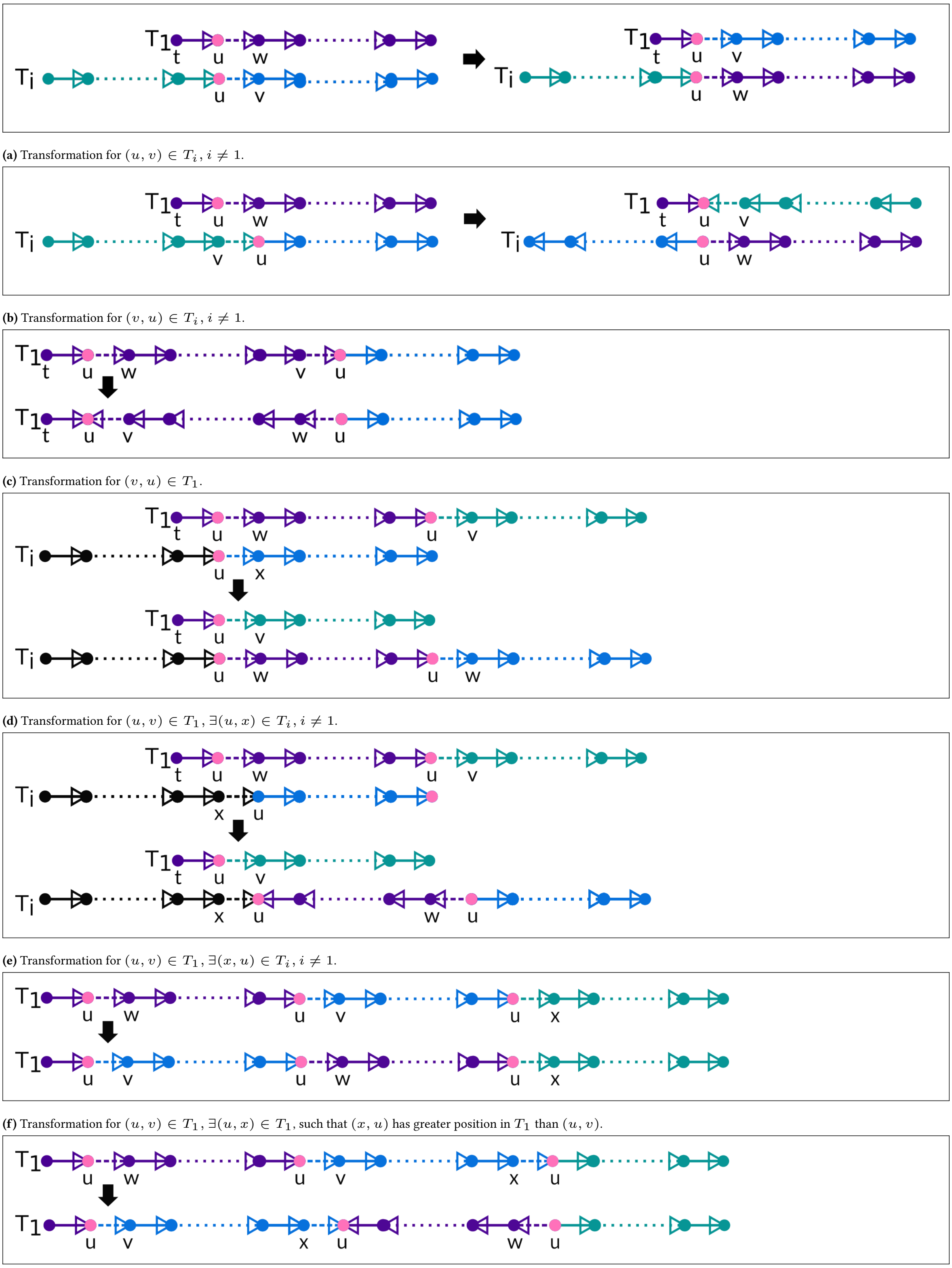
Lemma 4. *The connectivity certificate can be built in $O(log^2(n) \cdot m)$ amortized time and enables incrementing and decrementing the multiplicity of an edge in $O(log^2(n))$ amortized time.*

An efficient implementation of a *connectivity certificate* can be achieved by using the data structure based on Euler tour trees proposed by Holm *et al.* (2001). This data structure enables queries for the connectivity of two vertices in a graph, as well as the addition and removal s in amortized $O(log^2(n))$ time.

Since the KG graph is a multi-graph, an additional data structure is needed to store, for every vertex, a list of vertices incident to it along with their multiplicities. This can be implemented using AVL trees (Adelson-Velskii and Landis, 1962), which hich allow for the addition, removal, and querying for entries, as well as locating the next entry in $O(log(m))$ time, where m is the number of entries.

The dynamic graph can be built by adding all edges of the ALDKG to initially empty data structure proposed by Holm *et al.* (2001). This can be done in $O(log^2(n) \cdot m)$ amortized time, since the number of edges not incident to supplemental vertices is at most m and the number of edges incident to supplemental vertices does not exceed $2 \cdot m$. The dictionary storing all vertices and edges incident to them, along with their multiplicities, can be built by enumerating all edges of ALDKG in $O(\log(m) \cdot m)$ time.

Decrementing the multiplicity of an edge requires decrementing the corresponding multiplicities in the dictionary. If the resulting multiplicity is equal to 0, the edge should



(a) Transformation for $(u, v) \in T_i$, $i \neq 1$.

(b) Transformation for $(v, u) \in T_i$, $i \neq 1$.

(c) Transformation for $(v, u) \in T_1$.

(d) Transformation for $(u, v) \in T_1$, $\exists (u, x) \in T_i$, $i \neq 1$.

(e) Transformation for $(u, v) \in T_1$, $\exists (x, u) \in T_i$, $i \neq 1$.

(f) Transformation for $(u, v) \in T_1$, $\exists (u, x) \in T_1$, such that (x, u) has greater position in T_1 than (u, v).

(g) Transformation for $(u, v) \in T_1$, $\exists (x, u) \in T_1$, such that (x, u) has greater position in T_1 than (u, v).

Figure 4.3: Transformation of the *witness certificate* showing that the vertex v is a *good neighbor* of vertex u, given certain locations of other edges containing u. Segmental edges are depicted as solid lines, while adjacency edges are depicted as dashed lines. Arrows indicate the ordering of vertices in the edges which is preserved during transformations.

be removed from the dynamic graph, and the appropriate entries should be deleted from the dictionary. Similarly, incrementing the multiplicity of an edge requires incrementing the corresponding multiplicities in the dictionary if such entries exist. Otherwise, the edge should be added to the dynamic graph, and the corresponding entries should be added to the dictionary.

4.1.6. Traversal of the recursion tree

The *connectivity certificate* is employed to:

1. ensure the existence of at least one *good neighbor* of the currently processed vertex u (there are no dead ends in the recursion tree).
2. locate next *good neighbor* of the currently processed vertex u and save space by keeping only one in the recursion stack instead of a the list of vertices.
3. quickly skip nodes with only one *good neighbor* of u.

At least one good neighbor of u*.* Throughout the traversal of a recursion tree the invariant is that the *connectivity certificate* is connected, which guarantees, by Lemma 3, the existence of at least one MOED that extends the currently processed trail $s^T \rightsquigarrow u$. This, in turn, ensures that at least one *good neighbor* exists for the processed vertex u.

Next good neighbor of u*.* The next *good neighbor* is retrieved from the *vertex index* using ordering based on their *vertex numbers.* This ensures a reduction in the space of the recursion stack and guarantees that the produced MOEDs will be returned in lexicographic order, where the alphabet is a set of *edge representations* of trails.

Lemma 5. *Given a trail* $s^T \rightsquigarrow w = (s^T \rightsquigarrow u) \cdot (u, v) \cdot (v, w)$ *and a connectivity certificate* C *corresponding to it, the next good neighbor (in terms of ordering using vertex indices) can be found in* $O(log^2(n))$ *amortized time.*

Proof. First, multiplicities of edges (u, v) and (v, w) are increased by one in C. Then, the next entry after (u, v), such that it forms adjacency edge is retrieved from the *edge index* and verified if it is a *good neighbor* using C. If the answer is positive, the vertex is returned. Otherwise, the next vertex incident to u that form with u an adjacency edge is returned, as only one vertex can be the *uncertain vertex.* All operations concerning C take $O(log^2(n))$ amortized time.

Lemma 6. *Given the current node of the recursion tree and connectivity certificate* C *for its trail* $s^T \rightsquigarrow u$*, the certificates for the parent or for a child of the node can be computed in* $O(log^2(n))$ *amortized time.*

Proof. Let $(s^T \rightsquigarrow u')$ be a trail in the parent node, where v' is a good neighbor of u'. Then the connectivity certificate for a child can be obtained by decrementing the multiplicity of edges (u', v') and (v', w') by one in C.

Let $s^T \rightsquigarrow w = (s^T \rightsquigarrow u) \cdot (u, v) \cdot (v, w)$ be a C in the child node. The connectivity certificate for the parent node is obtained by increasing the multiplicity of edges (u, v) and (v, w) by one in C. All these operations take $O(log^2(n))$ amortized time.

Theorem 1. *Given an ALDKG all its MOEDs can be listed in lexicographic order (where the alphabet is a set of edge representations of trails) without duplicates with $O(log^2(n) \cdot l)$ time cost per solution. The initial setup time is $O(log^2(n) \cdot m)$, and the space complexity required for the algorithm is $O(log(m) \cdot l)$.*

By the definition of a *good neighbor* given a partial trail $s^T \rightsquigarrow u$ a simple induction shows that the algorithm will generate all MOEDs having $s^T \rightsquigarrow u$ as a prefix. No MOEDs will be listed twice for the same reason. Suppose, by contradiction, that there exists a MOED that is not outputted. Let $s^T \rightsquigarrow u$ be the recursion node that fails to generate a child. By the definition of a *good neighbor*, this is a contradiction, as using it we find all v that can extend to obtain $s^T \rightsquigarrow u \cdot (u, v)$

Let m_1, and m_2 be MOEDs and let $s^T \rightsquigarrow u$ be their common longest prefix sequence of edges in the trail, followed by the next edge. Let $s^T \rightsquigarrow u \cdot (u, v_1)$ and $s^T \rightsquigarrow u \cdot (u, v_2)$ be the respective trails, where $v_1 \neq v_2$. Due to the fact that prefixes of MOEDs are always extended by a pair of edges and the second one is a segmental edge determined unambiguously based on the first adjacency edge, the $\{u, v_1\}$ and $\{u, v_2\}$ are adjacency edges. Without loss of generality let us assume that $v_1 < v_2$. The recursion traverses child elements of a parent in an order of their *vertex number* so m_1 will be generated before m_2, ensuring that all MOEDs will be returned in lexicographic order (where the alphabet is a set of edge representations of trails).

Since the algorithm spends $O(log^2(n))$ amortized time in each node of the recursion tree, and there are l nodes for each solution, we get $O(log^2(n) \cdot l)$ amortized time per solution. The space complexity is $O(log(m) \cdot l)$, as the recursion stack can be at most $O(l)$ long and the memory per node is constant, AVL tree used in $O(m)$, and dynamic graph proposed by Holm *et al.* (2001) is $O(log(m) \cdot m)$.

4.2. Results and discussion

We have applied our enumeration algorithm for listing rearrangement scenarios for the case P5513_206 from Nazaryan-Petersen *et al.* (2018) study. Our algorithm found two additional scenarios supplementing the three plausible end products of rearrangement pre-

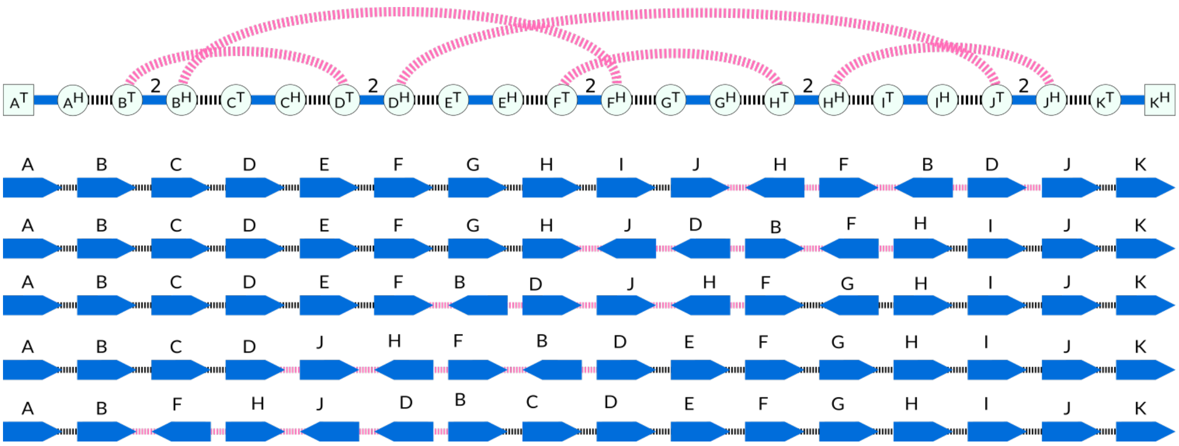

A^T A^H B^T 2 B^H C^T C^H D^T 2 D^H E^T E^H F^T 2 F^H G^T G^H H^T 2 H^H I^T I^H J^T 2 J^H K^T K^H
A B C D E F G H I J H F B D J K
A B C D E F G H J D B F H I J K
A B C D E F B D J H F G H I J K
A B C D J H F B D E F G H I J K
A B F H J D B C D E F G H I J K

Figure 4.4: Karyotype Graph and corresponding five rearranged genomes for case P5513_206 from the Nazaryan-Petersen *et al.* (2018) study.

sented in this study. Karyotype Graph modeling this CCR and corresponding five rearranged genomes are shown in Fig. 4.4.

The algorithm presented in this chapter can be further developed to enable listing of Minimal Eulerian Decomposition containing circular *derivative chromosomes*. One possible direction for extending the algorithm is to incorporate the notion of a centromere and enumerate *derivative chromosomes* with a sound number of such elements, or specify the complexity class of a problem defined in this manner.

The further directions for the development and clinical application of the presented algorithm include the incorporation of optical genome mapping data in the enumeration procedure. This enhancement would allow for pruning the recursion tree or assigning an alignment score to determine the quality of the match between a given rearrangement scenario and optical genome mapping data. Possible clinical applications also involve the analysis of highly rearranged cancer genomes.

5

`TADeus2`: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3D chromatin structure

> *"The greatest value of a picture*
> *is when it forces us to notice*
> *what we never expected to see."*
>
> — John Wilder Tukey

In previous chapters, we have described methods for disentangling structural rearrangements and shown that they have a significant impact on the course of evolution and serve as important sources of genetic variability. These rearrangements alter the chromatin structure, and can result in the misregulation of spatiotemporal gene expression, leading to disease, even if its breakpoints do not affect coding regions. In this chapter, we present a web server facilitating the diagnosis of the clinical consequences of such structural alterations.

In the last decades, the field of clinical genetics has been revolutionised by NGS technologies (Zhang *et al.*, 2011). One of the most popular approaches for identifying the molecular etiology of human diseases is the sequencing of coding regions, which constitute less than 3% of the genome (Hangauer *et al.*, 2013). However, the application of whole exome sequencing (WES) characterized the genotype-phenotype correlation in only 25–40% of patients (Sawyer *et al.*, 2016). This diagnostic rate can be attributed to the limitation of WES, as it targets only coding regions, while other causes of pathology, such as structural variants (SVs) and copy number variants (CNVs) are often located in non-coding DNA (Kumar *et al.*, 2021a; Zare *et al.*, 2017).

SV and CNV might lead to disease through 4 main mechanisms: (i) direct disruption or deletion of a gene leading to haploinsufficiency (Harewood *et al.*, 2012); (ii) generation of SV/CNV derived fusion genes – a mechanism often present, but not limited to cancer cells (Mitelman *et al.*, 2007; Mertens *et al.*, 2015; Eykelenboom *et al.*, 2012); (iii) changes in the gene dosage due to duplication (Lupski, 1999) and (iv) disruption of the epigenetic equilibrium caused by displacement of regulatory elements (Harewood *et al.*, 2010; Harewood and Fraser, 2014). The latter mechanism is often caused by changes in the three-dimensional (3D) chromatin structure. The 3D genome architecture controls spatiotemporal gene expression and plays a key role in the development and disease (Aigner *et al.*, 1974). At the (sub)megabase scale, the 3D chromatin structure is organized into topologically associated domains (TADs), delimited by boundaries enriched in CTCF binding sites (Dixon *et al.*, 2012; Chen *et al.*, 2019). TADs facilitate enhancer-promoter contacts within their bodies and insulate inter-TADs chromatin interactions (Flavahan *et al.*, 2019; Lupiáñez *et al.*, 2015).

Interestingly, the disruption of TADs or the formation of a novel TADs can lead to gene deregulation associated with congenital disorders e.g. brachydactyly, a limb malformation affecting finger development (Lupiáñez *et al.*, 2015), and Cooks syndrome, a congenital disorder affecting digits and nails (Franke *et al.*, 2016).

Despite the characterization of the genotype-phenotype correlation in a number of diseases linked to SVs and CNVs in non-coding DNA, the evaluation of their pathogenicity caused by the disruption of long-range regulatory interactions remains challenging (Lettice *et al.*, 2011; Harewood and Fraser, 2014). More evidence points towards the fact that computational investigation of non-coding variants severity is helpful in the diagnosis of genetic disorders (Wells *et al.*, 2019; Zhang and Lupski, 2015; Momozawa and Mizukami, 2020).

For this reason, there is a clear need for an easy to use web server that enables a quick evaluation of chromatin conformation changes and provides a visual framework for the interpretation of SVs/CNVs affecting TADs structures.

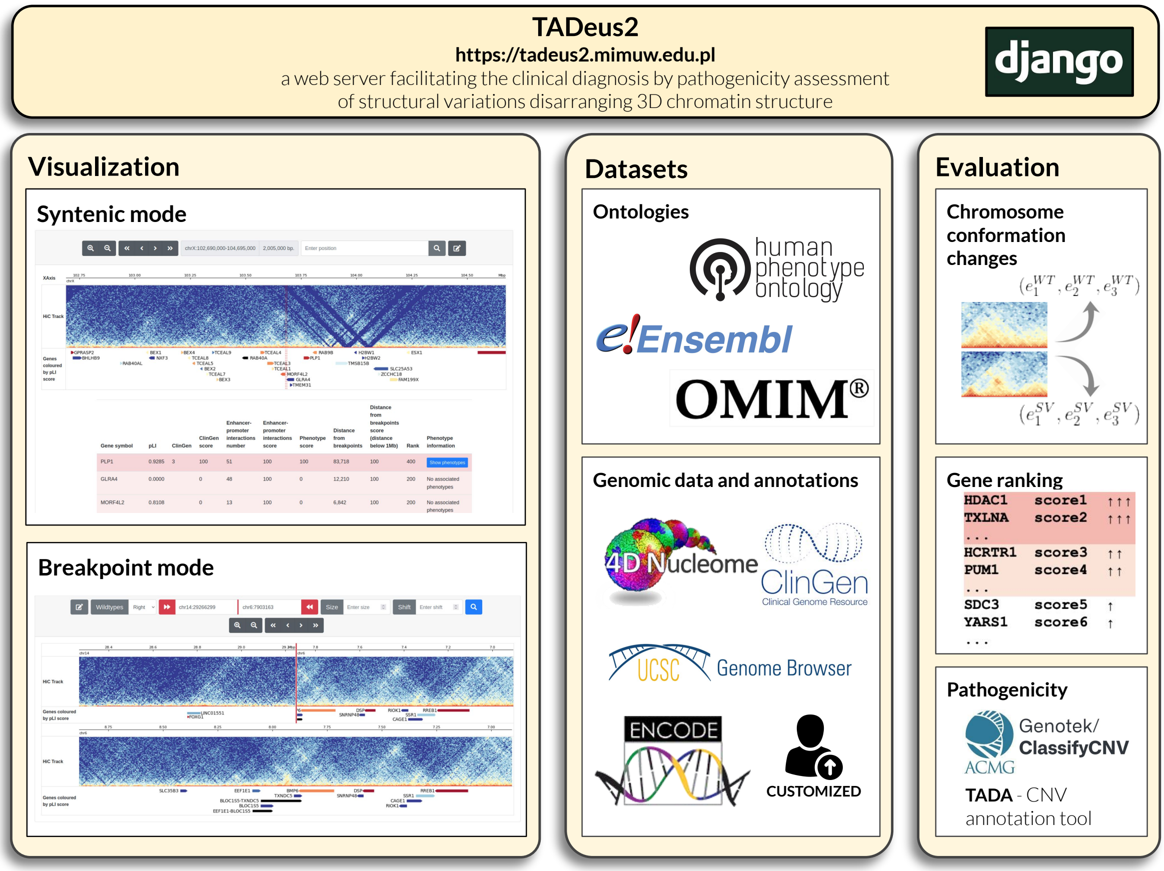


Figure 5.1: Main functionalities offered by TADeus2 together with their integration within the clinical workflow for the pathogenicity assessment of structural variations disarranging 3D chromatin structure.

5.1. Existing tools for the clinical evaluation of SV

The growing number of NGS data from symptomatic SVs/CNVs carriers led to the emergence of multiple databases inspiring further studies regarding the role of gene position effect (GPE) in the patient phenotype. For example, Zepeda-Mendoza *et al.* (2017) applied the haploinsufficiency and triplosensivity scores to characterize GPE derived from balanced translocations in 17 subjects from the Developmental Genome Anatomy Project (GDAP). Ruifeng et. al. using data from ClinVar database (Landrum *et al.*, 2017) introduced the Structure Influence score to prioritize and designate SVs that are likely to disturb gene regulation through TADs disorganization. Ibn-Salem *et al.* (2014) based on the DECIPHER database (containing nearly one thousand deletions) (Firth *et al.*, 2009) showed that only 4.5% of analyzed deletions can disrupt TAD boundaries leading to the gene misexpression due to enhancer adoption.

Investigation of NGS data from symptomatic patients carrying SVs led also to the development of multiple tools designed for clinical evaluation of SVs, including: 3Disease Browser (Li *et al.*, 2016), PhenogramViz (Köhler *et al.*, 2014), GeCCO (Hehir-Kwa *et al.*, 2010), or rankings based on HI scores (Huang *et al.*, 2010). However, PhenogramViz, GeCCO and

HI scores do not use the chromatin conformation and regulatory data for the SV-induced pathogenecity estimation, while 3Disease Browser does not evaluate any novel SVs.

Among the newest web-services providing an extended palette of features for SVs/CNVs evaluation in the clinical context are 3D Genome Browser (Wang *et al.*, 2018) and CNVxplorer (Requena *et al.*, 2021). The main functionalities provided by 3D Genome Browser consist of analysis of disease-associated SVs and 3D chromatin structure by visualization and integration of Hi-C and ChIP-seq data. This tool also enables inspection of inter-chromosomal interactions. However, it is limited only to the uni-directed strand view (hg19), and does not provide any views for comparisons of the rearranged genome structure vs. wild-type. The CNVxplorer mines a comprehensive set of clinical, genomic, and epigenomic features associated with CNVs making it one of the most versatile diagnosis tools available online. Nonetheless, it focuses only on CNVs and does not consider balanced rearrangements events. Summarized evaluation of tools with purpose comparable to TADeus2 is presented in the Table 5.1.

5.2. Our approach

To address existing challenges and provide a competitive tool, we have developed TADeus2 – a web server for a quick evaluation of SVs/CNVs that provides a visual framework to aid the medical expert in the interpretation of variant pathogenicity in the context of changes in the TADs organization (Fig. 5.1). Based on the type of a variant TADeus2 allows to visualize the affected region in two modes: (i) *syntenic* mode dedicated for the analysis of deletions and duplications; (ii) *breakpoint* mode designed for translocations and inversions. The former mode allows visualization of one genomic region, while the latter enables analysis of two different genome loci joined together during the rearrangement event. Both modes integrate multiple datasets (e.g. Hi-C or ChIPseq) either available on the server or provided by the user. The tool is user-friendly and allows to perform a customized analysis by providing the genomic coordinates of the analyzed variants. It should be emphasized that TADeus2 and its previous version TADeus (Poszewiecka *et al.*, 2018) was successfully used in the recently published studies Pienkowski *et al.* (2019, 2020).

5.3. Methods of clinical evaluation of SVs

Evaluation of the gene pathogenicity

To assess and rank the pathogenicity of a gene TADeus2 $score$ is introduced. For a given gene g its value is calculated based on the following indicators: (i) CG(g): the ClinGen haploinsufficiency/triplosensitivity score (Rehm *et al.*, 2015), (ii) EPdis(g): the number of

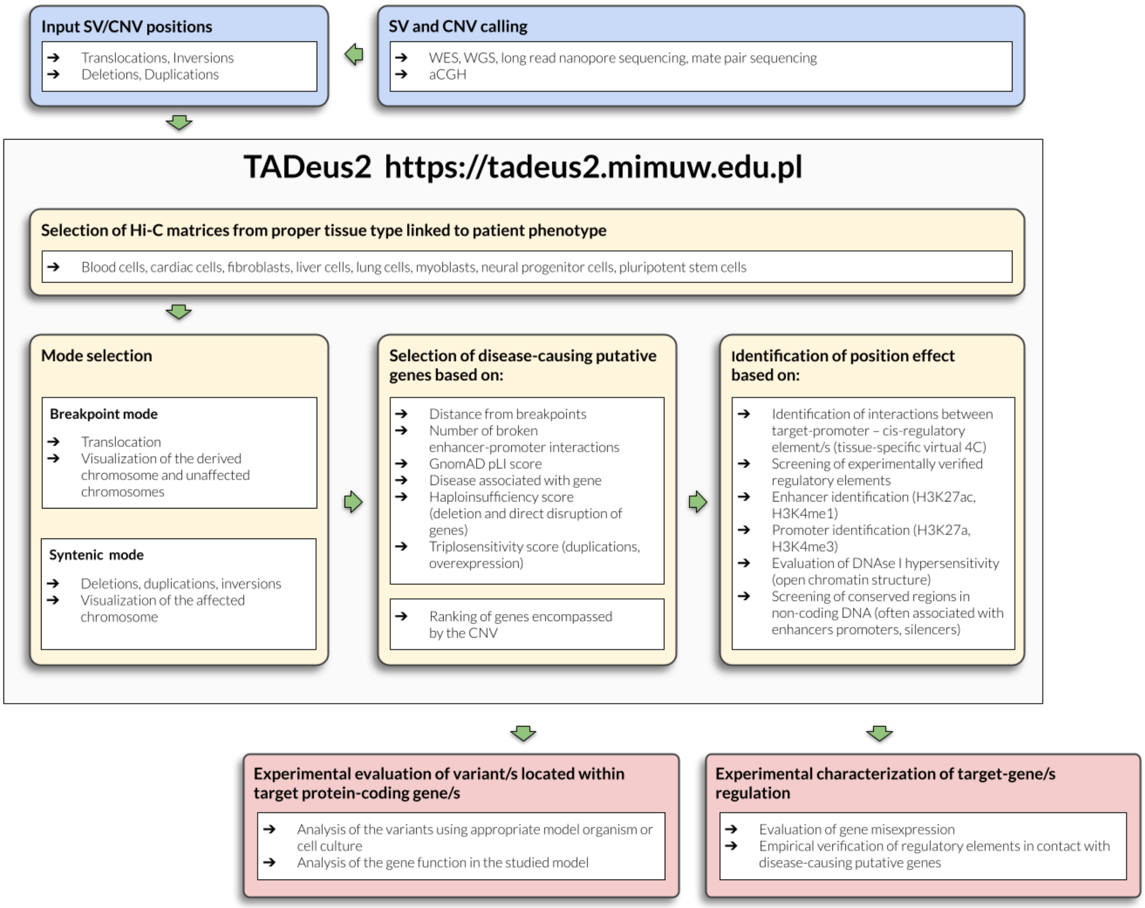

Input SV/CNV positions
Translocations, Inversions
Deletions, Duplications
SV and CNV calling
WES, WGS, long read nanopore sequencing, mate pair sequencing
aCGH
TADeus2 https://tadeus2.mimuw.edu.pl
Selection of Hi-C matrices from proper tissue type linked to patient phenotype
Blood cells, cardiac cells, fibroblasts, liver cells, lung cells, myoblasts, neural progenitor cells, pluripotent stem cells
Mode selection
Breakpoint mode
Translocation
Visualization of the derived chromosome and unaffected chromosomes
Syntenic mode
Deletions, duplications, inversions
Visualization of the affected chromosome
Selection of disease-causing putative genes based on:
Distance from breakpoints
Number of broken enhancer-promoter interactions
GnomAD pLI score
Disease associated with gene
Haploinsufficiency score (deletion and direct disruption of genes)
Triplosensitivity score (duplications, overexpression)
Ranking of genes encompassed by the CNV
Identification of position effect based on:
Identification of interactions between target-promoter – cis-regulatory element/s (tissue-specific virtual 4C)
Screening of experimentally verified regulatory elements
Enhancer identification (H3K27ac, H3K4me1)
Promoter identification (H3K27a, H3K4me3)
Evaluation of DNAse I hypersensitivity (open chromatin structure)
Screening of conserved regions in non-coding DNA (often associated with enhancers promoters, silencers)
Experimental evaluation of variant/s located within target protein-coding gene/s
Analysis of the variants using appropriate model organism or cell culture
Analysis of the gene function in the studied model
Experimental characterization of target-gene/s regulation
Evaluation of gene misexpression
Empirical verification of regulatory elements in contact with disease-causing putative genes

Figure 5.2: Overview of the clinical diagnosis workflow of `TADeus2` The above diagram outlines the subsequent steps recommended for the SVs/CNVs evaluation. The pipeline starts with the specification of a pair of coordinates accompanied by a Hi-C matrix of interest. Using two different views the structural rearrangements can be visualized and inspected. Next, based on numerous indicators of disease-causing putative genes candidates for further analysis can be selected. Finally, position effects can be evaluated thanks to provided databases from UCSC e.g. H3K27ac or experimentally verified regulatory elements. All the obtained outcomes are candidates for any further confirmation with wet experiments.

distant candidate enhancer–promoter predicted interactions disrupted by the breakpoints based on Thurman *et al.* (2012), (iii) HPO(g): the number of names and links to the associated phenotype described in Human Phenotype Ontology (HPO) – the ontology of phenotypic abnormalities with associated diseases and genes (Köhler *et al.*, 2016), (iv) dist(g): the distance from the rearrangement breakpoints. The final formula is accordingly a scaled sum of the four compounds:

$$\begin{aligned} score(g) =& \mathbb{I}\,(\,\mathrm{CG}(g) = 1\,) + \mathbb{I}\,(\,\mathrm{EPdis}(g) > 0\,) + \\ &+ \mathbb{I}\,(\,\mathrm{HPO}(g) > 0\,) + \mathbb{I}\,(\,\mathrm{dist}(g) < 1\mathrm{Mb}\,)\,, \end{aligned}$$

where CG(g) is equal to 1 when g falls into one of the ClinGen categories: Sufficient Evidence, Emerging Evidence, Autosomal Recessive; and $\mathbb{I}(c) = 1$, when a condition c is met and 0 otherwise.

As an example consider gene $g =$ TBR1 ranked as first in evaluation of structural variant involving chromosome 2 (exemplary case study from the main page of the web service). This gene is annotated in ClinGen, so $\mathrm{CG}(g) = 1$, there are 38 disrupted distant enhancer-promoter interactions, i.e. $\mathrm{EPdis}(g) > 0$, also the $\mathrm{HPO}(g) > 0$ as several abnormal phenotypes have been found in HPO, and finally the distance of the TBR1 gene from the breakpoint is small enough, i.e. $\mathrm{dist}(g) < 1\mathrm{Mb}$. In total we get a Total Pathogenecity Score equal to 4.

`TADeus2` rank gene table

The table is ordered by the aforementioned ranking score in the descending order. Genes with the equal scores are secondary sorted by the number of the disrupted enhancer–promoter interactions and the distance from the rearrangement breakpoints. For the convenience of the user the final importance ranking is color-coded with a dark-pink to white scale. Additionally, for a broader perspective, the disease and inheritance type from Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005) and pLI score from Genome Aggregation Database (gnomAD) (Karczewski *et al.*, 2020), were added.

SVs/CNVs pathogenicity assessment

To evaluate the pathogenicity of a rearrangement `TADeus2` uses the state-of-the-art, third-party software: TADA (available only for autosomes) (Hertzberg *et al.*, 2022) and ClassifyCNV (Gurbich and Ilinsky, 2020) as well as an original statistical significance p-value.

TADA automatically ranks Copy Number Variants (CNVs) based on a extensive catalogue of functional annotations supported by enrichment analysis. The software is based on a machine-learning classifier to accurately predict and prioritize pathogenic deletion or duplication to produce a well-calibrated pathogenicity score.

ClassifyCNV, uses pre-parsed publicly available databases to calculate a pathogenicity score for each duplication and deletion. Importantly, the tool is an implementation of the 2019 ACMG guidelines for variant interpretation that provide a set of criteria to score variants and place them into one of the five classification tiers (Riggs *et al.*, 2020).

Finally, we propose calculation of an empirical p-value to assess the statistical significance of the predicted number of interrupted interactions between the *cis*-regulatory elements, which is a good predictor of SVs pathogenicity (D'haene and Vergult, 2021).

To this end we compute the null hypothesis probability distribution - the probability distribution for all possible numbers of disrupted interactions induced by a random breakpoint. We sample 10^4 loci and compute the number of potential disruptions that would occur because of a breakpoint in that loci 5.3. The resulting distribution is approximated by a mixture of the 0-concentrated Dirac distribution and the geometric distribution:

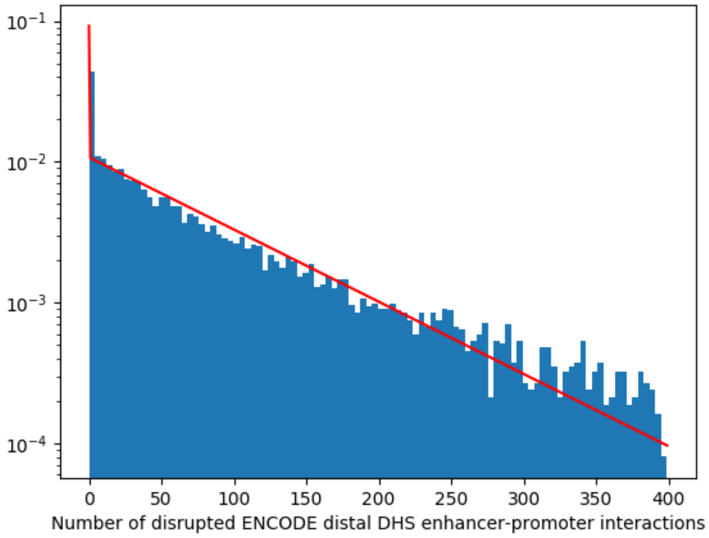


Figure 5.3: Distribution of number of disrupted distal enhancer–promoter interactions induced by a random breakpoint. Histogram of the number of distal enhancer-promoter interactions for 10000 randomly chosen breakpoint loci (blue bars) and the probability density function given by equation 5.3 with parameters estimated using this sample (red line).

$$\mathbb{P}(X = k|p, \delta) = \begin{cases} \delta, & \text{for } k = 0 \\ (1-\delta)p(1-p)^{k-1}, & \text{for } k > 0 \\ 0, & \text{for } k < 0 \end{cases}$$

for some $p \in (0, 1)$ and $\delta \in (0, 1)$ parameters, which we estimated from the data (Fig. 5.3). The proportion of breakpoints in all genome loci that do not disrupt any predicted promoter–enhancer interactions is an estimator of δ (0.0924), while p (0.01174) was estimated as the maximum likelihood estimation (MLE). As a result, the smallest number of enhancer–promoter interactions broken by a rearrangement breakpoint that is statistically significant (i.e p-value ≤ 0.05) is equal to 246.

5.4. Diagnosis workflow of TADeus2 web server

The input of TADeus2 consists of breakpoint coordinates (hg38) previously identified experimentally. The analysis starts with the selection of an adequate Hi-C matrix (currently, data from 8 different cell-types are available). The Hi-C matrix is displayed as *a triangle plot* (upper triangle rotated by 45 degrees) and TADs structure is visualized (red lines), in order to investigate the SV role in reorganization of the domains (often linked to symptomatic gene misexpression). Based on the type of a variant the user should choose either *syntenic* mode (deletion or duplication), or *breakpoint* mode (translocation, inversion). The former, presents a continuous sequence of a genomic region and is designed to generate a ranked table of putative disease-causing genes. The latter integrates two genomic fragments from different loci (e.g. located at distinct chromosomes) as if it was a continuous genome chunk, allowing visualization of the new rearrangement. To use this mode the user should specify the coordinates of the breakpoints and the direction of the fused strands (e.g. in the case of translocation between p and q arms, a forward reverse strand fusion occurs). The breakpoint view also includes wild-type regions that compose the rearrangement. Further analysis (in both modes) includes an integration of additional tracks containing publicly available or self-uploaded datasets. After the selection of disease causing putative genes (based on the TADeus2rank gene table or the interpretation of particular tracks e.g. gnomAD pLI) the user should focus on the characterization of the cis-regulatory landscape within the region of interest. This can be achieved by: (i) identification of candidate enhancers (CEs) based on the histone marks (H3K27ac, H3K4me1), chromatin accessibility (DNAseI), and conservation; (ii) investigation of the track with experimentally validated enhancers; (iii) analysis of putative enhancer-promoter interactions based on Virtual 4C. Finally, verification of the regulatory elements, enhancer-promoter interactions, disrupted genes should be further confirmed experimentally. For the convenience of the reader, the overview of the clinical diagnosis workflow is also presented in Fig. 5.2.

5.5. Validation of TADeus2 gene ranking scheme

Comparison of TADeus2ranking schemes with ranking scheme presented in Mendoza *et al.*

Zepeda-Mendoza *et al.* (2017) analyzed 17 subjects with apparently balanced chromosomal abnormalities with breakpoints in the non-coding regions (15 translocations and 2 inversions). Using their second ranking scheme, they predicted 116 genes for exhibiting position effects (15 genes as strong candidates and 101 genes as probable candidate). Second ranking scheme will be used to validate TADeus2 ranking scheme that prioritize genes near rearrangement breakpoint according to their clinical relevance.

TADeus2 ranking scheme was validated by analyzing chromosomal rearrangement presented in Zepeda-Mendoza *et al.* (2017). The method implemented in TADeus2 predicted 156 genes that may exhibit position effects (25 genes as strong candidates and 131 genes as probable candidates). TADeus2 ranking and second ranking scheme from Zepeda-Mendoza *et al.* (2017) predicted 113 common genes. Fig. 5.4 presents Venn diagram showing sets of genes predicted as strong and probable candidates for exhibiting position effects by Zapeda-Mendoza second ranking scheme and the TADeus2 ranking scheme.

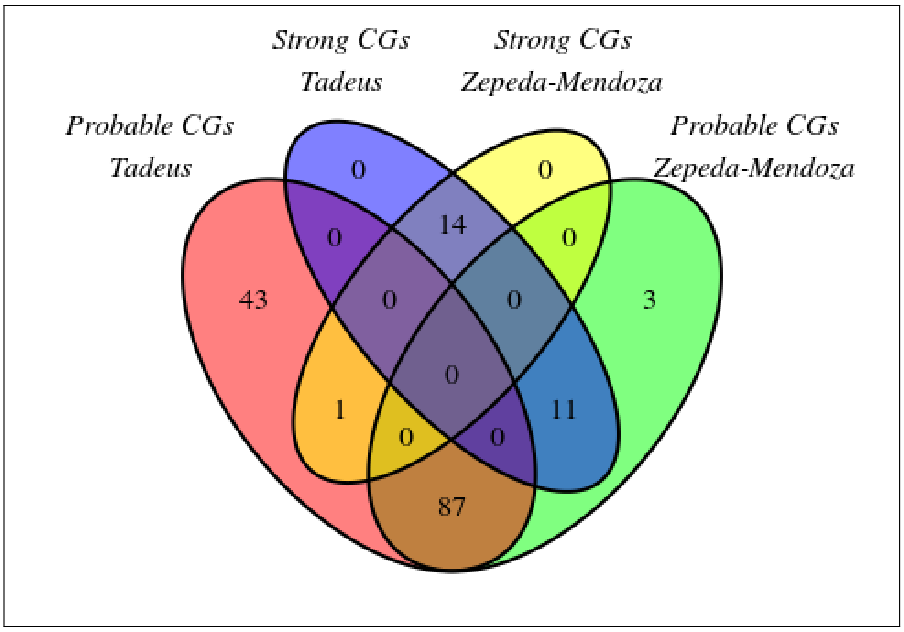


Figure 5.4: Venn diagram showing sets of candidate genes (CGs) predicted for exhibiting position effects by Zapeda-Mendoza second ranking scheme and `TADeus2` ranking scheme. Genes predicted by Zepeda-Mendoza second ranking scheme as strong candidates are shown in yellow while probable candidates are shown in green. Genes predicted by `TADeus2` ranking scheme as strong candidates are shown in blue while probable candidates are show in red.

Validation of `TADeus2` gene ranking scheme on data obtained from the literature

In order to test `TADeus2` accuracy a detailed search of literature was performed to find well-described cases of position effects generated by SVs and CNVs. For each case, breakpoints, clinical presentation and genes that according to authors contribute to the disease were collected. A table ranking disease-causing putative genes close to the SV/CNV was generated only for the chromosome where the disease causing gene was present. In total, 21 distinct cases were found and used to validate the `TADeus2` ranking method. For 18 (85,7%) cases, the genes contributing to the disease were predicted as the strongest candidates (top scores if compared to the rest of the genes localized on the same chromosome). In the remaining 3 (14,3%) cases, genes were predicted as the most probable candidates with lower scores.

5.6. Use-cases of `TADeus2` workflow in a clinical diagnosis setting

Below we present in detail three selected cases from the set described above an one case of position-effect in patient from Baylor College of Medicine clinical chromosomal microarray database of 65,000 patient.

Case 1 – Position effect in a balanced translocation neighbouring *FOXG1* gene.

Murcia Pienkowski *et al.* identified the exact structure of a balanced chromosomal translocation (46,XX,t(6;14)(p25.1;q12)) in a patient with epileptic seizures and severe developmental delay (Pienkowski *et al.*, 2019). The breakpoint on chromosome 6 disrupted the*TXNDC5* gene in the second intron (chr6:7903201 hg38), while chromosome 14 was damaged in non-coding DNA (chr14:29266318). TADeus2 ranking was used to establish genes that have a higher probability of being responsible for the disease. On chromosome 6 the highest score was assigned to gene *FDSP*. At the same time on chromosome 14 *FOXG1* was chosen as the gene with the highest probability to impact the phenotype.

Based on the clinical picture that the patient displayed we decided to conduct the next steps of the analysis using data Hi-C from neural progenitor cells (NPC). The following tracks have been added to the breakpoint view: Genes coloured by pLI score, Virtual 4C for promoter regions (+/- 2.5 kb from transcription start site (TSS)) for *DSP* and *FOXG1* and experimentally validated functional elements from NCBI database (Fig. 5.5). The result indicate that *FOXG1* lost contact with 4 experimentally verified enhancers, while *DSP* did not lose any, making *FOXG1* the most probable cause of the disease (see the red box in Fig. 5.5). Indeed, at least 11 translocations located in proximity of *FOXG1* in patients with Rett-like syndrome have been described. Furthermore, it has been proposed that loss of only one active *FOXG1* enhancer is not clinically relevant (Mehrjouy *et al.*, 2018). This is consistent with our results as the analyzed translocation leads to displacement of two apparently important enhancers: hs433 and hs342; and two enhancers with unknown impact on *FOXG1*: hs1168 and hs598. hs433 is thought to bring other enhancers into physical contact with *FOXG1* (Ibn-Salem *et al.*, 2014), while hs342 is the main candidate for the regulation of *FOXG1* expression (Mehrjouy *et al.*, 2018). Overall, TADeus2 indicated correctly the gene responsible for the disease as well as important regulatory elements that might have been at the root of the patient phenotype. Importantly, further experimental investigation of cis-regulatory landscape encompassing *FOXG1* might shed more light into the etiology of the patients phenotype.

Case 2 – Position effect in an inversion accompanied with a deletion located nearby *DLX5* and *DLX6* genes.

Kerry Brown et al. identified a paracentric inversion of the long arm of chromosome 7, inv(7)(q21.3q35) in five patients with hearing loss and craniofacial defects (Brown *et al.*, 2010). The breakpoint in q21.3 (accompanied by a 5.1 kb deletion (chr7:96,935,329-96,940,443; hg38)) is located within non-coding region, 65-80 kb away from *DLX5*and *DLX6* genes,

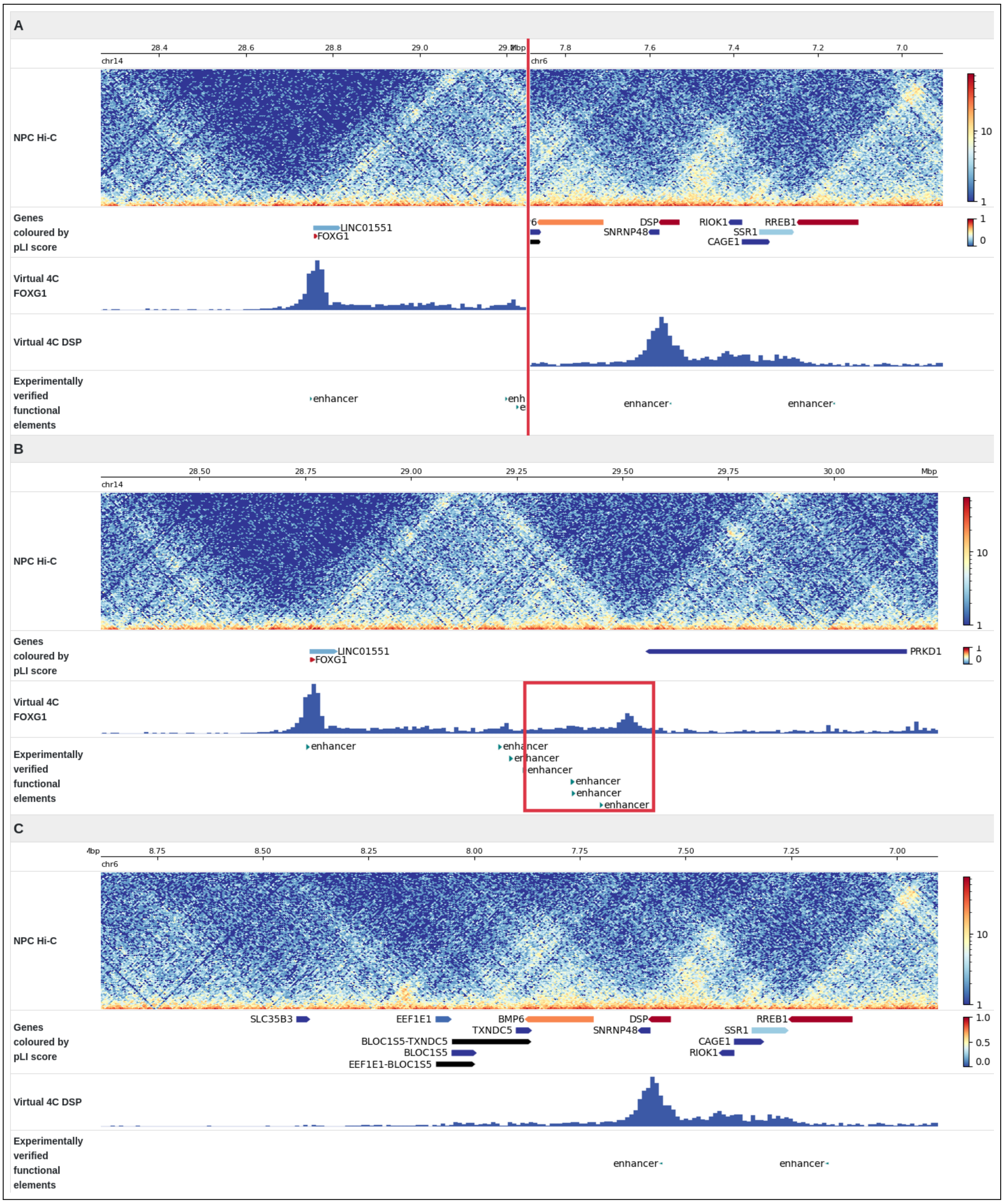


Figure 5.5: Analysis of a balanced translocation 46,XX,t(6;14)(p25.1;q12) in a patient with epileptic seizures and severe developmental delay. Visualization of the breakpoint fusion chromosome der(14) (A) and wild type chromosomes: chromosome 14 (B) chromosome 6 (C). NPC Hi-C data was used to generate the chromatin architecture (top track). The figure contains the following tracks: gene gnomAD pLI score (bottom track) in color-scale (0=blue and 1=red), virtual 4C for promoter regions (+/- 2.5 kb from TSS) for FOXG1 and DSP and experimentally validated functional elements. The enhancers displaced by the translocation are marked with the red rectangle.

while the 7q35 breakpoint disrupts the *CNTNAP2* gene, between exon two and ten.

First, we used TADeus2 to analyze the potential role of 7q35 breakpoint in the molecular etiology of hearing loss and craniofacial defects. Characterization of the chromatin architecture surrounding the breakpoint region revealed no genes within the TAD except *CNTNAP2* (Fig. 5.6a). However, the pLI score for *CNTNAP2* was equal to zero, suggesting that disruption of one copy of *CNTNAP2* has a low probability of being responsible for the abovementioned defects. Indeed, the spatiotemporal expression analysis of *CNTNAP2* during mouse embryonic development revealed no expression in the majority of tissues linked to the patient's phenotype (Brown *et al.*, 2010). Moreover, the *CNTNAP2*-/- mice present no abnormalities in craniofacial and inner ear development (Poliak *et al.*, 2003). Therefore, for further analysis we focused on the 7q21.3 breakpoint region.

The `TADeus2` ranking list pointed *DLX5* and *DLX6* as genes with a high probability of being responsible for these developmental abnormalities (top two genes with rank score: 300), with the pLI score equal to 0.2184 and 0.9213, respectively. Based on these results we extended our analysis by implementing the `TADeus2` browser tracks with: experimentally validated functional elements from NCBI database and virtual 4C for *DLX5* and *DLX6* promoter regions (+/- 2.5 kb from TSS). Multiple enhancers are located within the genomic region affected by the inversion, and in close 3D proximity with *DLX5* and *DLX6* promoters (based on virtual 4C).

This suggests that the inversion might disrupt long-range contacts between six experimentally validated enhancers (in the red rectangle) and promoters of *DLX5* and *DLX6*, most likely leading to a misregulation of these two genes (Fig. 5.6b)).

Importantly, *DLX5* is expressed in the otic placode and vesicle (involved in formation of inner ear vestibular structures), and in the semicircular canals of the inner ear (Merlo *et al.*, 2002b). Deletion of *DLX5* and *DLX6* leads to dysplastic ears and congenital deafness (Chromosome 7 Annotation Project, http://www.chr7.org), which is consistent with the mouse phenotype. The *DLX5* knockout as well as *DLX5* and *DLX6* double knockout mice are characterized by the craniofacial malformations and abnormalities in the ear development (Merlo *et al.*, 2002b; Acampora *et al.*, 1999; Merlo *et al.*, 2002a). Collectively, `TADeus2` results align with literature results suggesting that changes in the *DLX5* and *DLX6* gene dosage caused by disruption of enhancer-promoter connections may lead to the patient phenotype.

Case 3 - Position effect on *PLP1* may cause a subset of Pelizaeus-Merzbacher disease symptoms

Pelizaeus-Merzbacher disease (PMD, MIM 312080) is an X linked recessive dysmyelination disorder in which formation of meylin in the central nervous system is affected. PMD typically manifests with nystagmus, spastic quadriplegia, ataxia, and developmental delay. Mutation within the *PLP1* gene are responsible for 15–20% of PMD cases. This gene is triplosensitive and its duplications accounts for the majority of PMD cases (60–70%).

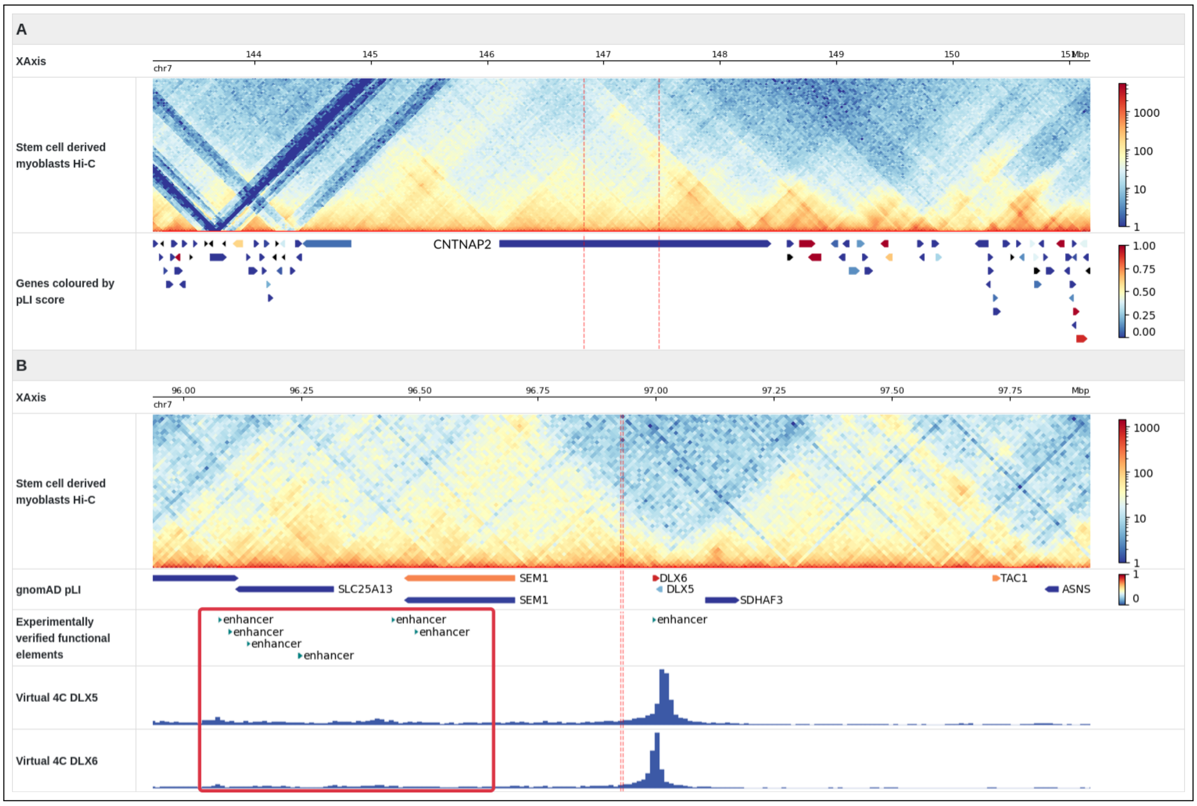


Figure 5.6: Analysis of the inversion inv(7)(q21.3q35) linked to the craniofacial defects and hearing loss. (A) Visualisation of the 7q35 breakpoint located between *CNTNAP2* exon two and ten. Stem cell derived myoblasts were used to visualise the chromatin architecture (top track) combined with gene gnomAD pLI score (bottom track) in color-scale (0=blue and 1=red). (B) Visualisation of the breakpoint in q21.3 located centromeric to *DLX5* and *DLX6*. The Hi-C data was used from stem cell derived myoblasts (top track), followed by analysis of gene gnomAD pLI scores, and also cis-regulatory elements (middle tracks). The putative interactions between *DLX5*, *DLX6* promoters (+/- 2.5 kb from TSS) and non-coding DNA were investigated using Virtual 4C (bottom track). The enhancers affected by inversion are marked with the red rectangle.

Therefore, position effect may by possible cause PMD in the remaining 10–25% of patients.

Muncke *et al.* (2004) reported a patient with a subset of the PMD symptoms, including moderate intellectual disability and cerebellar ataxia associated with dysmyelination. Patient and his unaffected mother carried an apparently balanced inversion with the breakpoints in Xp22 and Xq22. The breakpoint on the short arm of the X chromosome does not disrupt any gene and is unlikely to be causative as patient's mother is unaffected. Thus, the breakpoint on the long arm the chromosome X is probably responsible for the pathogenesis as it cannot be compensated for in the male patient.

Physical mapping revealed that Xq22 breakpoint disrupts a putative pseudogene *GLRA4* and resides 70 Kb upstream to *PLP1*. The genomic region in the neighborhood of the breakpoint is shown in Fig. 5.7. Muncke *et al.* (2004) excluded *GLRA4* as a source of the dysmyelination defect and suggested that position effect on *PLP1* is causative for a subset of PMD symptoms in the affected patient.

TADeus2 ranking was used to find the gene responsible for this pathogenesis. It included 100 genes in the 3 Mb regions flanking the Xq22 breakpoint. *PLP1* was the only

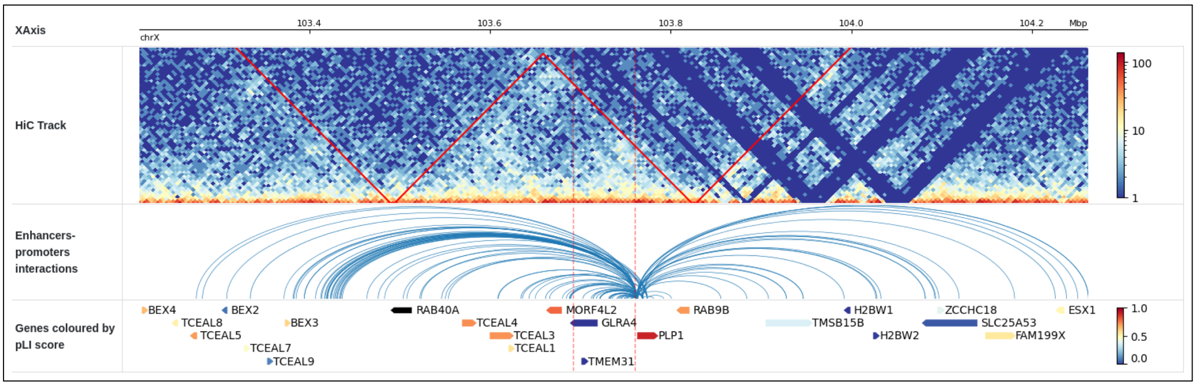


Figure 5.7: Analysis of an inversion 46, XY, inv(X) (p22.3; q22) in a patient suffering from a subset of PMD symptoms including moderate mental retardation. Visualization of ± 0.5 Mb region flanking inversion breakpoint on Xq22 region. The precise breakpoint site is located in the 70 Kb region downstream PLP1 and is depicted by two vertical red lines. The figure contains the following tracks: (i) Hi-C data from GM12878 Human B-lymphoblastoids cell line, (ii) distant candidate enhancer–promoter interaction filtered to those associated with *PLP1* gene visualised as arcs, (iii) gene gnomAD pLI score in color-scale (0=blue and 1=red)

gene classified as a strong candidate for exhibiting position effect and 18 genes were predicted as probable candidates. *PLP1* has a Clingen score of 3, indicating a sufficient evidence for dosage pathogenicity. The rearrangement breakpoint disrupts 55 predicted enhancer–promotes interactions, the distance from the rearrangement breakpoints is smaller than 1 Mb and has the assigned phenotypes in OMIM and HPO, thus *PLP1* gets the highest possible score (4). The list of phenotypes associated with *PLP1* displayed in the ranking table includes Pelizaeus-Merzbacher disease.

Case 4 - Position effect on *SLC7A7* in patient with *de novo* microdeletion 14q11.2

We queried the Baylor College of Medicine clinical chromosomal microarray database of 65,000 patients and identified 11 CNVs in nine patients that do not overlap any protein-coding sequence.

Patient with familial thrombocytopenia and leukemia in Chromosomal Microarray Analysis (CMA) revealed a deletion within chromosome band 14q11.2 spanning approximately 0.448 Mb. The presence of smaller 0.04 Mb interval within the 0.448 Mb loss indicated a homozygous deletion. Fig. 5.8 shows a ± 0.8 Mb neighborhood of that SVs.

TADeus2 has been used to rank the genes near the breakpoints of this rearrangement according to their potential for exhibiting position effects. *SLC7A7* is the highest ranked gene and it gets the highest possible score as it fulfills all the criteria used in the ranking scheme for gene evaluation: its ClinGen haploinsufficiency score is 30 indicating association with autosomal recessive phenotype, the rearrangement breakpoints disrupt 28 putative enhancer-promoter interactions, the distance from the gene to the rearrangement breakpoint is 340,742 bases, and the gene has associated phenotypes in OMIM and HPO.

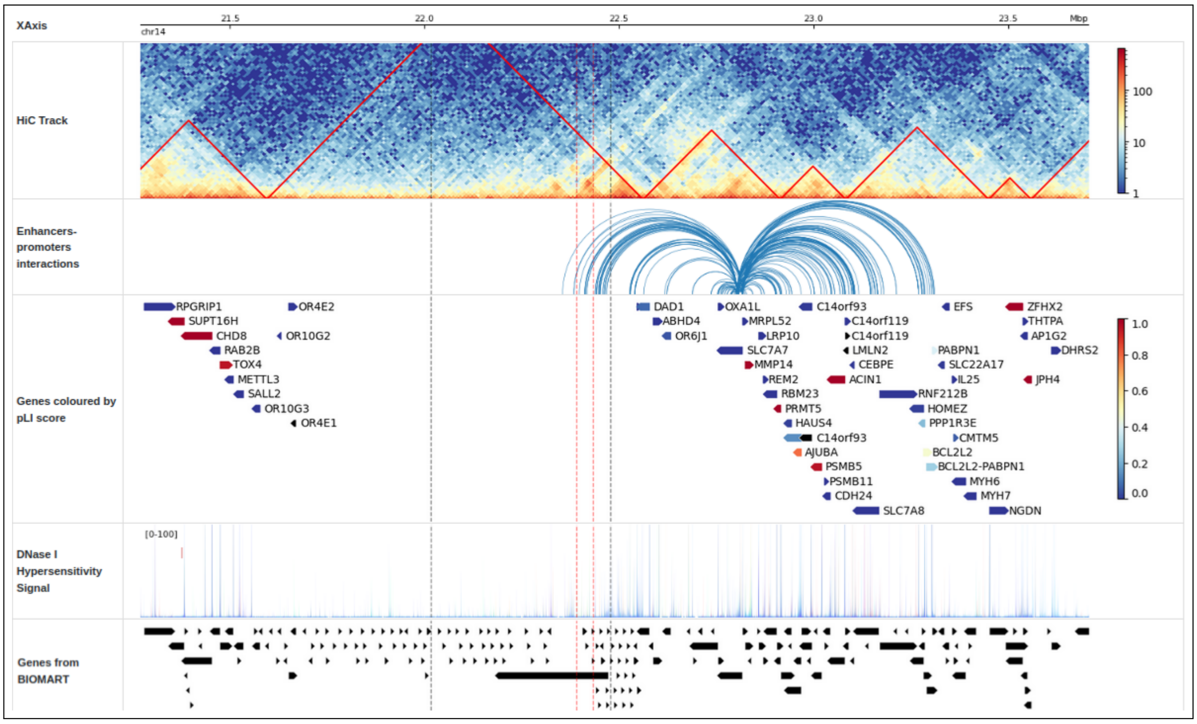


Figure 5.8: Analysis of *de novo* microdeletion del(14)(q11.2) in a patient suffering from familial thrombocytopenia and leukemia Visualization of region flanking deletions located centromeric to *SLC7A7* gene. The breakpoints of 0.448 Mb deletion are visualised as vertical black lines and smaller internal 0.04 Mb deletion as red lines. The figure contains the following tracks: (i) Hi-C data from GM12878 Human B-lymphoblastoids cell line, (ii) distant candidate enhancer–promoter interaction filtered to those associated with *SLC7A7* gene visualised as arcs, (iii) gene gnomAD pLI score in color-scale (0=blue and 1=red), (iv) DNase I Hypersensitivity Signal from Encode, (v) genes from BIOMART database.

Table 5.2 shows genes in 3 Mb neighborhood of deletion that are predicted as strong or probable candidates for exhibiting position effects.

SLC7A7 is associated with lysinuric protein intolerance (LPI, MIM #222700) that is caused by homozygous or compound heterozygous mutation of that gene. LPI is associated with a variety of clinical symptoms, including hematologic abnormalities like thrombocytopenia. One of the phenotypes in HPO associated with *SLC7A7* is thrombocytopenia and it consistent with patient phenotype.

It should be noted that smaller homozygous deletion includes the T-cell receptor alpha gene cluster. Heterozygous copy number changes in this region are common polymorphic findings. Clinical significance of this subregion, which includes the TRDC (T-cell receptor delta chain C region) is currently not clear. The DECIPHER patient (338650) with a similar homozygous deletion has T-cell acute lymphoblastic leukemia.

5.7. TADeus2 web server implementation and functionality

TADeus2 can serve as a genome browser that display Hi-C matrices in conjunction with diverse genomic assays outputs. The user can create a plot and add tracks using the preloaded public data or its own datasets. Currently, TADeus2 allows users to upload the genomic data in BED (BED3, BED6, BED9 and BED12) and BEDGraph format.

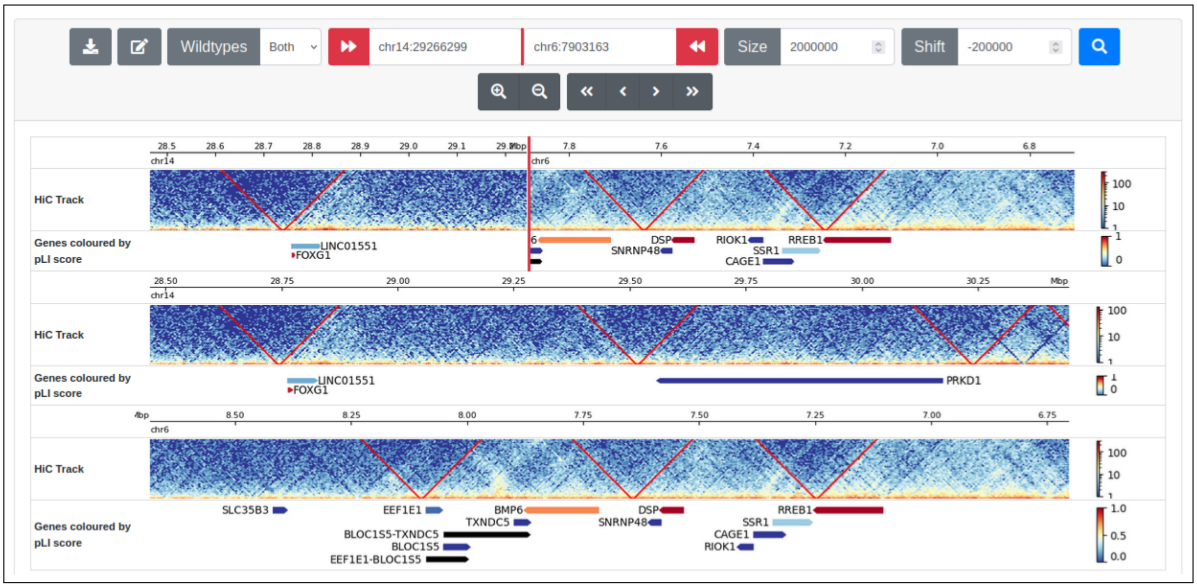
Intrachromosomal interaction matrices are visualized with the upper triangle rotated by 45 degrees with bins aligned to the corresponding chromosomal coordinates. Additionally TADs can be plotted as triangles on such rotated heatmap. TADeus2 can visualize one dimensional discrete genomic features, such as genes, enhances or SVs as tiles. Data representing genes (in BED12 format) can be displayed in flybase format or in format that shows introns. Data from 3C, ChIA-Pet experiments or enhancer-promoter pairs can be visualized as arcs whereas domains as triangles. To emphasize some important genomic loci, the user can provide the coordinates that will be marked by the vertical dashed lines.

Breakpoint mode - innovative genome browser functionality

TADeus2 genome browser provides innovative breakpoint mode for visualizing region flanking rearrangement locus as continuous fragment (Fig. 5.9). Additionally, in this mode, wild-type regions can be displayed to show the entire view of regions that were fused by the rearrangement enabling exploring genomic features that were perturbed by the rearrangement. Intuitive interface enables setting the coordinates and direction of fused regions, resize the visualized region, and shift the rearrangement breakpoint from the center of the plot to either side. Notably, to date, there is no genome browser with such functionality (Table 5.1).

Implementation and architecture of TADeus2 web server

TADeus2 is implemented as a Django application using MySQL database as data storage. Fragments of code from HiCExplorer (Ramírez *et al.*, 2018) are reused in the track plot module. All the presented command-line tools (e.g TADA, classifyCNV) incorporated in TADeus2 are accompanied with a responsive and user-friendly graphical interface developed with Bootstrap v4. Additionally, *JavaScript* snippets enable customization of forms and export of any user-defined plots that are produced using the python *matplotlib* package. Finally, TADeus2 provides the rest-API that can be utilized by external applications.



Wildtypes Both

chr14:29266299

chr6:7903163

Size 2000000

Shift -200000

Figure 5.9: User interface of breakpoint mode of TADeus2 genome browser. Figure presents the user interface of the genome browser in breakpoint mode. At the top, a panel for defining the coordinates and directions of fused regions together with the properties of visualized regions is shown. Widgets from panel enable resizing the shown region, shifting the rearrangement breakpoint from the center of the plot, and customizing the wild-type options. Below, three tracks showing genomic data from the region defined using the user interface are shown. The top track presents a fused rearranged region. As the "shift" number field from the user interface panel is set to the negative value (-200,000), the breakpoint (red vertical line) is moved to the left side, so the region of the right fused strand takes slightly more space than the left one. The middle and bottom tracks present entire wild-type regions enabling the user to inspect their genomic features extended beyond the rearrangement breakpoint.

Available experimental data

Currently, TADeus2 allows the user to upload genomic data in BED and BEDGraph format. The user is also provided with over 30 preloaded, publicly available datasets described in detail at https://tadeus2.mimuw.edu.pl/datasources/.

Web server instance and code availability

TADeus2 is publicly available at https://tadeus2.mimuw.edu.pl. Detailed tutorial, help pages, short videos presenting functionality of application, and a FAQ section are provided. The tool is free and open to all users and there is no login requirement. TADeus2 source code and installation instructions are available at https://github.com/bposzewiecka/tadeus2. The software is distributed under a GNU General Public License v3.0.

5.8. Conclusions

In this chapter, we presented TADeus2 web server with all its features. The tool is an online solution for visualization and analysis of SVs and CNVs in the context of chromatin conformation. Thanks to an easy to use and user-friendly interface, TADeus2 is suitable as a handful framework for preliminary clinical diagnosis of patients for non-bioinformatician experts.

The future directions for the improvement of the web server include additional expansions of the breakpoint view that would allow the view of more breakpoint coordinates at once (particularly useful for visualisation of inversions, duplications, and complex rearrangements i.e. chromothripsis). In addition, we plan to implement a functionality that provides a visualization of chromatin conformation in 3D.

Moreover, to predict the TAD structure affected by the SVs, Akita software (Fudenberg *et al.*, 2020) will be used to generate a new Hi-C matrix for the rearranged TAD. This functionality is currently under development, with a beta version available for the user. Furthermore, the statistical modeling of chromatin organization disorders caused by structural variants will be further developed to achieve higher accuracy in its assessment. Lastly, we state that `TADeus2` will benefit from regular updates using the wealth of data publicly available.

	Options					Breakpoint browser options			
Name of service	HiC Heat map	Tad annotation	Virtual 4C view	Data upload or local installation	Breakpoint browser mode	Wildtype view	Customised tracks	Unlimited number of HiC tracks	Head/Tail orientation of coordinates
`TADeus2`	✓	✓	✓	✓	✓	✓	✓	✓	✓
3D Genome Browser[1] (Wang *et al.*, 2018)	✓	✓	✓	✓	✓		✓[2]		
3DIV[3] (Yang *et al.*, 2018)	✓	✓	✓	✓	✓				
HiGlass (Kerpedjiev *et al.*, 2018)	✓	✓	✓	✓					
Juicebox (Durand *et al.*, 2016)	✓		✓	✓					
WashU Epigenome Browser (Li *et al.*, 2019)	✓			✓					
HUGIN (Martin *et al.*, 2017)	✓	✓	✓						
3Disease Browser (Li *et al.*, 2016)	✓	✓							

Table 5.1: Listing of all available tools that provide similar functionalities to `TADeus2` regarding typical clinical use cases, TAD analysis and breakpoint viewing.

[1] Inter-chromosomal interaction mode is regarded as breakpoint mode for comparisons.

[2] As UCSC or WashU session link.

[3] Complex SV and 3D genome view is regarded as breakpoint mode for comparisons.

Table 5.2: Ranking of genes within 3 Mb neighborhood of rearrangement breakpoint 14q11.2(22901689-22942482) predicted as strong or probable candidates for exhibiting position effect

Gene symbol	pLI	ClinGen	Enhancer–promoter interaction number	Distance from breakpoints	Phenotypes in OMIM or HPO	Rank
SLC7A7	0.0021	30	28	340,742	Yes	4
CEBPE	0.0024	30	0	646,343	Yes	3
MRPL52	0.0001	NA	12	361,764	No	2
LRP10	0.0002	NA	10	408,307	No	2
PRMT5	0.9948	NA	8	456,312	No	2
OXA1L	0.0000	NA	5	298,525	No	2
DAD1	0.2765	NA	3	115,693	No	2
ABHD4	0.0001	NA	2	138,783	No	2
OR6E1P	NA	NA	2	229,792	No	2
TRAC	NA	NA	0	78,615	Yes	2
MMP14	0.9871	NA	0	375,754	Yes	2
PABPN1	0.5998	NA	0	852,912	Yes	2
SALL2	0.0000	NA	0	912,457	Yes	2
MYH6	0.0000	NA	0	935,004	Yes	2
MYH7	0.0001	0	0	962,445	Yes	2
RPGRIP1	0.0000	30	0	1,123,022	Yes	2
TGM1	0.0000	30	0	1,791,156	Yes	2
PNP	0.0656	30	0	1,964,576	Yes	2

6

Conclusions and future research

"Zniknął za rogiem
I przepadł jak szyszka
Ale nie płaczmy, bo
Ale nie płaczmy, bo
Nie o to chodzi by złowić króliczka,
Ale by gonić go,
Ale by gonić go,
Ale by gonić go!"

— Agnieszka Osiecka ,"Króliczek"

This dissertation encompasses various approaches to investigate genome architecture, primarily focusing on the detection and interpretation of chromosomal rearrangements while also exploring their contribution to evolutionary history.

A significant contribution of this research is the introduction of an innovative approach for the local assembly of regions enriched in segmental duplications. This algorithm follows a novel bottom-up paradigm for constructing contigs. The successful application of this tool in reconstructing selected chimpanzee subtelomeres has led to the formulation of a valuable hypothesis concerning the impact of the ancestral fusion event on human evolution. Furthermore, an improved method for estimating the temporal scope of major evolutionary events was introduced. This method, along with its new application in estimating speciation events, offers enhanced accuracy in understanding the timeline of evolutionary processes. Additionally, a procedure for enumerating all possible scenarios of complex chromosomal rearrangements was presented. This method proved valuable in the analysis of structural alterations in a patient harboring such rearrangements, providing insights into the underlying mechanisms of their formation. Lastly, a web service was developed to facilitate the clinical evaluation of changes in genome architecture caused by structural variants. This software introduces a novel genome browser that provides a unique visualization perspective on genomic features from the vantage point of rearrangement breakpoints. Overall, this dissertation presents a comprehensive exploration of genome architecture, unveiling new methodologies and tools that contribute to our understanding of chromosomal rearrangements and their significance in both evolutionary and clinical contexts.

Based on the work presented in this dissertation and the interdisciplinary nature of this research, there are several potential avenues for future exploration that can contribute to both the field of algorithmics and clinical applications.

Firstly, future work could involve refining and expanding the algorithmic approaches for detecting and interpreting chromosomal rearrangements. By incorporating advanced machine learning techniques and leveraging large-scale genomic datasets, we can aim to enhance the accuracy and efficiency of structural variant analysis. Such advancements would not only benefit the algorithmics field but also have direct implications for clinical applications, improving the detection and understanding of genomic changes associated with structural variants in patients and their impact on phenotypic traits.

Furthermore, there is an opportunity to explore the evolutionary implications of chromosomal rearrangements in a broader range of species. By studying the impact of rearrangements in diverse organisms and their subsequent effects on speciation events, we can gain a deeper understanding of the evolutionary dynamics underlying genomic architecture. Combining genomics and evolutionary biology can provide valuable insights into the role of rearrangements in shaping species divergence and adaptation.

Additionally, the development of the web service introduced in this dissertation opens up avenues for further improvement and expansion. Future work could involve incorporating additional features into the service, such as predictive modeling of disease risk based on structural variants. Integrating population-level data into the analysis can provide a more comprehensive understanding of the prevalence and impact of rearrangements in different populations. Furthermore, the inclusion of interpretability tools can aid clinicians in effectively comprehending and communicating the implications of genomic changes associated with rearrangements, ultimately benefiting clinical decision-making and patient care.

By pursuing these directions for future research, the interdisciplinary nature of this work can continue to yield significant contributions to both algorithmics and clinical applications. The integration of advanced algorithms, evolutionary insights, and clinical relevance holds the potential to revolutionize personalized medicine, improve diagnostic capabilities, and enhance patient outcomes.

Bibliography

Acampora, D., Merlo, G. R., Paleari, L., Zerega, B., Postiglione, M. P., Mantero, S., Bober, E., Barbieri, O., Simeone, A. and Levi, G. (1999). Craniofacial, vestibular and bone defects in mice lacking the distal-less-related gene dlx5. *Development*, **126** (17), 3795–3809.

Adelson-Velskii, G. M. and Landis, E. M. (1962). An algorithm for organization of information. In *Doklady Akademii Nauk*, Russian Academy of Sciences, vol. 146, pp. 263–266.

Aganezov, S. and Raphael, B. J. (2020). Reconstruction of clone-and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Genome research*, **30** (9), 1274–1290.

—, Zban, I., Aksenov, V., Alexeev, N. and Schatz, M. C. (2019). Recovering rearranged cancer chromosomes from karyotype graphs. *BMC bioinformatics*, **20** (20), 1–11.

Aigner, A., Müller, G., Knapp, E. and Raas, E. (1974). Systolic time intervals, phonocardiograms and sound spectrograms in patients with starr-edwards aortic valve prostheses. *Cardiology*, **59** (1), 30–40.

Allshire, R. C., Gosden, J. R., Cross, S. H., Cranston, G., Rout, D., Sugawara, N., Szostak, J. W., Fantes, P. A. and Hastie, N. D. (1988). Telomeric repeat from t. thermophila cross hybridizes with human telomeres. *Nature*, **332** (6165), 656–659.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21** (1).

Avarello, R., Pedicini, A., Caiulo, A., Zuffardi, O. and Fraccaro, M. (1992). Evidence for an ancestral alphoid domain on the long arm of human chromosome 2. *Human genetics*, **89** (2), 247–249.

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M. *et al.* (2013). Punctuated evolution of prostate cancer genomes. *Cell*, **153** (3), 666–677.

Baldini, A., Ried, T., Shridhar, V., Ogura, K., D'Aiuto, L., Rocchi, M. and Ward, D. C. (1993). An alphoid dna sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Human genetics*, **90** (6), 577–583.

Birmelé, E., Ferreira, R., Grossi, R., Marino, A., Pisanti, N., Rizzi, R. and Sacomoto, G. (2013). Optimal listing of cycles and st-paths in undirected graphs. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, SIAM, pp. 1884–1896.

Brown, K. K., Reiss, J. A., Crow, K., Ferguson, H. L., Kelly, C., Fritzsch, B. and Morton, C. C. (2010). Deletion of an enhancer near dlx5 and dlx6 in a family with hearing loss, craniofacial defects, and an inv (7)(q21. 3q35). *Human genetics*, **127** (1), 19.

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., Meyer, T. J., Herrero, J., Roos, C., Aken, B., Anaclerio, F., Archidiacono, N., Baker, C., Barrell, D., Batzer, M. A., Beal, K., Blancher, A., Bohrson, C. L., Brameier, M., Campbell, M. S., Capozzi, O., Casola, C., Chiatante, G., Cree, A., Damert, A., de Jong, P. J., Dumas, L., Fernandez-Callejo, M., Flicek, P., Fuchs, N. V., Gut, I., Gut, M., Hahn, M. W., Hernandez-Rodriguez, J., Hillier, L. W., Hubley, R., Ianc, B., Izsvk, Z., Jablonski, N. G., Johnstone, L. M., Karimpour-Fard, A., Konkel, M. K., Kostka, D., Lazar, N. H., Lee, S. L., Lewis, L. R., Liu, Y., Locke, D. P., Mallick, S., Mendez, F. L., Muffato, M., Nazareth, L. V., Nevonen, K. A., O'Bleness, M., Ochis, C., Odom, D. T., Pollard, K. S., Quilez, J., Reich, D., Rocchi, M., Schumann, G. G., Searle, S., Sikela, J. M., Skollar, G., Smit, A., Sonmez, K., ten Hallers, B., Terhune, E., Thomas, G. W., Ullmer, B., Ventura, M., Walker, J. A., Wall, J. D., Walter, L., Ward, M. C., Wheelan, S. J., Whelan, C. W., White, S., Wilhelm, L. J., Woerner, A. E., Yandell, M., Zhu, B., Hammer, M. F., Marques-Bonet, T., Eichler, E. E., Fulton, L., Fronick, C., Muzny, D. M., Warren, W. C., Worley, K. C., Rogers, J., Wilson, R. K. and Gibbs, R. A. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, **513** (7517), 195–201.

Carvalho, C. and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, **17** (4), 224–238.

Chaisson, M. J., Mukherjee, S., Kannan, S. and Eichler, E. E. (2017). Resolving multicopy duplications de novo using polyploid phasing. In *International Conference on Research in Computational Molecular Biology*, Springer, pp. 117–133.

Chan, Y. C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C. C., Pei, K. J. and Vigilant,

L. (2010). Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. *PLoS ONE*, **5** (12), e14419.

Chatterjee, H. J., Ho, S. Y., Barnes, I. and Groves, C. (2009). Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.*, **9**, 259.

Chen, X., Ke, Y., Wu, K., Zhao, H., Sun, Y., Gao, L., Liu, Z., Zhang, J., Tao, W., Hou, Z. *et al.* (2019). Key role for ctcf in establishing chromatin structure in human embryos. *Nature*, **576** (7786), 306–310.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M. and Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, **437** (7055), 88–93.

Chiatante, G., Giannuzzi, G., Calabrese, F. M., Eichler, E. E. and Ventura, M. (2017). Centromere destiny in dicentric chromosomes: new insights from the evolution of human chromosome 2 ancestral centromeric region. *Molecular biology and evolution*, **34** (7), 1669–1681.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A. *et al.* (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, **13** (12), 1050–1054.

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., Glessner, J. T., Mason, T., Pregno, G., Dorrani, N. *et al.* (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome biology*, **18** (1), 1–21.

Costantini, M., Clay, O., Auletta, F. and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Res*, **16** (4), 536–541.

De Pagter, M. S., Van Roosmalen, M. J., Baas, A. F., Renkens, I., Duran, K. J., Van Binsbergen, E., Tavakoli-Yaraki, M., Hochstenbach, R., Van Der Veken, L. T., Cuppen, E. *et al.* (2015). Chromothripsis in healthy individuals affects multiple protein-coding genes and can result in severe congenital abnormalities in offspring. *The American Journal of Human Genetics*, **96** (4), 651–656.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002). Capturing chromosome conformation. *science*, **295** (5558), 1306–1311.

D'haene, E. and Vergult, S. (2021). Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. *Genetics in Medicine*, **23** (1), 34–46.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485** (7398), 376–380.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C. *et al.* (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, **16** (10), 1299–1309.

Dreszer, T. R., Wall, G. D., Haussler, D. and Pollard, K. S. (2007). Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome research*, **17** (10), 1420–1430.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S. and Aiden, E. L. (2016). Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell systems*, **3** (1), 99–101.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, **10**, 285–311.

Dutrillaux, B. (1979). Chromosomal evolution in primates: Tentative phylogeny from microcebus murinus (prosimian) to man. *Human Genetics*, **48** (3), 251–314.

Eisfeldt, J., Pettersson, M., Petri, A., Nilsson, D., Feuk, L. and Lindstrand, A. (2021). Hybrid sequencing resolves two germline ultra-complex chromosomal rearrangements consisting of 137 breakpoint junctions in a single carrier. *Human genetics*, **140** (5), 775–790.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred II error probabilities. *Genome Research*, **8** (3), 186–194.

Eykelenboom, J. E., Briggs, G. J., Bradshaw, N. J., Soares, D. C., Ogawa, F., Christie, S., Malavasi, E. L., Makedonopoulou, P., Mackie, S., Malloy, M. P. *et al.* (2012). A t (1; 11) translocation linked to schizophrenia and affective disorders gives rise to aberrant chimeric disc1 transcripts that encode structurally altered, deleterious mitochondrial proteins. *Human molecular genetics*, **21** (15), 3374–3386.

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. and Trask, B. J. (2002a). Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14. 1 and paralogous regions on other human chromosomes. *Genome research*, **12** (11), 1651–1662.

—, Newman, T., Linardopoulou, E. and Trask, B. J. (2002b). Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. *Genome Res.*, **12** (11).

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M. and Carter, N. P. (2009). Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, **84** (4), 524–533.

Flavahan, W. A., Drier, Y., Johnstone, S. E., Hemming, M. L., Tarjan, D. R., Hegazi, E., Shareef, S. J., Javed, N. M., Raut, C. P., Eschle, B. K. *et al.* (2019). Altered chromosomal topology drives oncogenic programs in sdh-deficient gists. *Nature*, **575** (7781), 229–233.

Flint, J., Wilkie, A. O., Buckle, V. J., Winter, R. M., Holland, A. J. and McDermid, H. E. (1995). The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nature genetics*, **9** (2), 132–140.

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L. *et al.* (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, **538** (7624), 265–269.

Fudenberg, G., Kelley, D. R. and Pollard, K. S. (2020). Predicting 3d genome folding from dna sequence with akita. *Nature methods*, **17** (11), 1111–1117.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Pr?fer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., H?ber, B., H?ffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, ., Gu?ic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science*, **328** (5979), 710–722.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.*, **43** (10), 1031–1034.

Grossi, R., Marino, A. and Versari, L. (2018). Efficient algorithms for listing k disjoint st-paths in graphs. In *LATIN 2018: Theoretical Informatics: 13th Latin American Symposium, Buenos Aires, Argentina, April 16-19, 2018, Proceedings 13*, Springer, pp. 544–557.

Gurbich, T. A. and Ilinsky, V. V. (2020). Classifycnv: a tool for clinical annotation of copy-number variants. *Scientific reports*, **10** (1), 1–7.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33** (suppl_1), D514–D517.

Han, J., Zhang, Z. and Wang, K. (2018). 3c and 3c-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, **11** (1), 1–10.

Hangauer, M. J., Vaughn, I. W. and McManus, M. T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding rnas. *PLoS genetics*, **9** (6), e1003569.

Harewood, L., Chaignat, E. and Reymond, A. (2012). Structural variation and its effect on expression. In *Genomic Structural Variants*, Springer, pp. 173–186.

— and Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Human Molecular Genetics*, **23** (R1), R76–R82.

—, Schütz, F., Boyle, S., Perry, P., Delorenzi, M., Bickmore, W. A. and Reymond, A. (2010). The effect of translocation-induced nuclear reorganization on gene expression. *Genome research*, **20** (5), 554–564.

Hehir-Kwa, J. Y., Wieskamp, N., Webber, C., Pfundt, R., Brunner, H. G., Gilissen, C., de Vries, B. B., Ponting, C. P. and Veltman, J. A. (2010). Accurate distinction of pathogenic from benign cnvs in mental retardation. *PLoS computational biology*, **6** (4), e1000752.

Hertzberg, J., Mundlos, S., Vingron, M. and Gallone, G. (2022). TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs. *Genome Biology*, **23** (1).

Hey, J. (2010). The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.*, **27** (4), 921–933.

Hillier, L. W., Graves, T. A., Fulton, R. S., Fulton, L. A., Pepin, K. H., Minx, P., Wagner-McPherson, C., Layman, D., Wylie, K., Sekhon, M., Becker, M. C., Fewell, G. A., Delehaunty, K. D., Miner, T. L., Nash, W. E., Kremitzki, C., Oddy, L., Du, H., Sun, H., Bradshaw-Cordum, H., Ali, J., Carter, J., Cordes, M., Harris, A., Isak, A., van Brunt, A., Nguyen, C., Du, F., Courtney, L., Kalicki, J., Ozersky, P., Abbott, S., Armstrong, J., Belter, E. A., Caruso, L., Cedroni, M., Cotton, M., Davidson, T., Desai, A., Elliott, G., Erb, T., Fronick, C., Gaige, T., Haakenson, W., Haglund, K., Holmes, A., Harkins, R., Kim, K., Kruchowski, S. S., Strong, C. M., Grewal, N., Goyea, E., Hou, S., Levy, A., Martinka, S., Mead, K., McLellan, M. D., Meyer, R., Randall-Maher, J., Tomlinson, C., Dauphin-Kohlberg, S., Kozlowicz-Reilly, A., Shah, N., Swearengen-Shahid, S., Snider, J., Strong, J. T., Thompson, J., Yoakum, M., Leonard, S., Pearman, C., Trani, L., Radionenko, M., Waligorski, J. E., Wang, C., Rock, S. M., Tin-Wollam, A. M., Maupin, R., Latreille, P., Wendl, M. C., Yang, S. P., Pohl, C., Wallis, J. W., Spieth, J., Bieri, T. A., Berkowicz, N., Nelson, J. O., Osborne, J., Ding, L., Meyer, R., Sabo, A., Shotland, Y., Sinha, P., Wohldmann, P. E., Cook, L. L., Hickenbotham, M. T., Eldred, J., Williams, D., Jones, T. A., She, X., Ciccarelli, F. D., Izaurralde, E., Taylor, J., Schmutz, J., Myers, R. M., Cox, D. R., Huang, X., McPherson, J. D., Mardis, E. R., Clifton, S. W., Warren, W. C., Chinwalla, A. T., Eddy, S. R., Marra, M. A., Ovcharenko, I., Furey, T. S., Miller, W., Eichler, E. E., Bork, P., Suyama, M., Torrents, D., Waterston, R. H. and Wilson, R. K. (2005). Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, **434** (7034), 724–731.

Hobolth, A., Christensen, O. F., Mailund, T. and Schierup, M. H. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.*, **3** (2), e7.

—, Dutheil, J. Y., Hawks, J., Schierup, M. H. and Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.*, **21** (3), 349–356.

Holm, J., De Lichtenberg, K. and Thorup, M. (2001). Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity. *Journal of the ACM (JACM)*, **48** (4), 723–760.

Hotaling, S., Kelley, J. L. and Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, **118** (52), e2109019118.

Huang, N., Lee, I., Marcotte, E. M. and Hurles, M. E. (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics*, **6** (10), e1001154.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach,

J. and Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, **24** (4), 688–696.

Ibn-Salem, J., Köhler, S., Love, M. I., Chung, H.-R., Huang, N., Hurles, M. E., Haendel, M., Washington, N. L., Smedley, D., Mungall, C. J. *et al.* (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome biology*, **15** (9), 423.

Ijdo, J., Baldini, A., Ward, D., Reeders, S. and Wells, R. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences*, **88** (20), 9051–9055.

il Sohn, J. and Nam, J.-W. (2016). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, p. bbw096.

Jauch, A., Wienberg, J., Stanyon, R., Arnold, N., Tofanelli, S., Ishida, T. and Cremer, T. (1992). Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences*, **89** (18), 8611–8615.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Salinas, C. A. A., Ahmad, T., Albert, C. M., Ardissino, D., Atzmon, G., Barnard, J., Beaugerie, L., Benjamin, E. J., Boehnke, M., Bonnycastle, L. L., Bottinger, E. P., Bowden, D. W., Bown, M. J., Chambers, J. C., Chan, J. C., Chasman, D., Cho, J., Chung, M. K., Cohen, B., Correa, A., Dabelea, D., Daly, M. J., Darbar, D., Duggirala, R., Dupuis, J., Ellinor, P. T., Elosua, R., Erdmann, J., Esko, T., Färkkilä, M., Florez, J., Franke, A., Getz, G., Glaser, B., Glatt, S. J., Goldstein, D., Gonzalez, C., Groop, L., Haiman, C., Hanis, C., Harms, M., Hiltunen, M., Holi, M. M., Hultman, C. M., Kallela, M., Kaprio, J., Kathiresan, S., Kim, B.-J., Kim, Y. J., Kirov, G., Kooner, J., Koskinen, S., Krumholz, H. M., Kugathasan, S., Kwak, S. H., Laakso, M., Lehtimäki, T., Loos, R. J. F., Lubitz, S. A., Ma, R. C. W., MacArthur, D. G., Marrugat, J., Mattila, K. M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Meigs, J. B., Melander, O., Metspalu, A., Neale, B. M., Nilsson, P. M., O'Donovan, M. C., Ongur, D., Orozco, L., Owen, M. J., Palmer, C. N. A., Palotie, A., Park, K. S., Pato, C., Pulver, A. E., Rahman, N., Remes, A. M., Rioux, J. D., Ripatti, S., Roden, D. M., Saleheen, D., Salomaa, V., Samani, N. J., Scharf, J., Schunkert, H., Shoemaker, M. B., Sklar, P., Soininen, H., Sokol, H., Spector, T., Sullivan, P. F., Suvisaari, J., Tai, E. S., Teo, Y. Y., Tiinamaija, T., Tsuang, M., Turner, D., Tusie-Luna, T., Vartiainen, E., Vawter, M. P., Ware, J. S., Watkins, H., Weersma, R. K., Wessman, M., Wilson, J. G., Xavier, R. J., Neale, B. M., Daly, M. J. and and, D. G. M. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature*, **581** (7809), 434–443.

Kasai, F., Takahashi, E., Koyama, K., Terao, K., Suto, Y., Tokunaga, K., Nakamura, Y. and Hirai, M. (2000). Comparative FISH mapping of the ancestral fusion point of human chromosome 2. *Chromosome Res.*, **8** (8), 727–735.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, **12** (6), 996–1006.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N. *et al.* (2018). Higlass: web-based visual exploration and analysis of genome interaction maps. *Genome biology*, **19** (1), 1–12.

Khrameeva, E., Kurochkin, I., Han, D., Guijarro, P., Kanton, S., Santel, M., Qian, Z., Rong, S., Mazin, P., Sabirov, M., Bulat, M., Efimova, O., Tkachev, A., Guo, S., Sherwood, C. C., Camp, J. G., Pääbo, S., Treutlein, B. and Khaitovich, P. (2020). Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Research*, **30** (5), 776–789.

Kloosterman, W. P., Guryev, V., van Roosmalen, M., Duran, K. J., de Bruijn, E., Bakker, S. C., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M. *et al.* (2011). Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human molecular genetics*, **20** (10), 1916–1924.

Köhler, S., Schoeneberg, U., Czeschik, J. C., Doelken, S. C., Hehir-Kwa, J. Y., Ibn-Salem, J., Mungall, C. J., Smedley, D., Haendel, M. A. and Robinson, P. N. (2014). Clinical interpretation of cnvs with cross-species phenotype data. *Journal of medical genetics*, **51** (11), 766–772.

—, Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M. *et al.* (2016). The human phenotype ontology in 2017. *Nucleic acids research*, **45** (D1), D865–D876.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, **37** (5), 540–546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, **27** (5), 722–736.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28** (19), 2520–2522.

Kumar, A., Adhikari, S., Kankainen, M. and Heckman, C. A. (2021a). Comparison of structural and short variants detected by linked-read and whole-exome sequencing in multiple myeloma. *Cancers*, **13** (6), 1212.

Kumar, V., Goutam, R. S., Umair, Z., Park, S., Lee, U. and Kim, J. (2021b). Foxd4l1. 1 negatively regulates chordin transcription in neuroectoderm of xenopus gastrula. *Cells*, **10** (10), 2779.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. *et al.* (2017). Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, **46** (D1), D1062–D1067.

Lejeune, J., Dutrillaux, B., Rethoré, M. O. and Prieur, M. (1973). Comparaison de la structure fine des chromatides d'homo sapiens et de pan troglodytes. *Chromosoma*, **43** (4), 423–444.

Lese, C. M., Fantes, J. A., Riethman, H. C. and Ledbetter, D. H. (1999). Characterization of physical gap sizes at human telomeres. *Genome research*, **9** (9), 888–894.

Lettice, L. A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P. S., Gautier, P., Morrison, H., Fantes, J., Hill, R. E. and FitzPatrick, D. R. (2011). Enhancer-adoption as a mechanism of human developmental disease. *Human mutation*, **32** (12), 1492–1499.

Li, D., Hsu, S., Purushotham, D., Sears, R. L. and Wang, T. (2019). Washu epigenome browser update 2019. *Nucleic acids research*, **47** (W1), W158–W165.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32** (14), 2103–2110.

— and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25** (14), 1754–1760.

Li, R., Liu, Y., Li, T. and Li, C. (2016). 3disease browser: a web server for integrating 3d genome and disease-associated chromosome rearrangement data. *Scientific reports*, **6**, 34651.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O. *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326** (5950), 289–293.

Liu, P., Erez, A., Nagamani, S. C. S., Dhar, S. U., Kołodziejska, K. E., Dharmadhikari, A. V., Cooper, M. L., Wiszniewska, J., Zhang, F., Withers, M. A. *et al.* (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, **146** (6), 889–903.

Luke, S. and Verma, R. S. (1995). Human (Homo sapiens) and chimpanzee (Pan troglodytes) share similar ancestral centromeric alpha satellite DNA sequences but other fractions of heterochromatin differ considerably. *Am. J. Phys. Anthropol.*, **96** (1), 63–71.

Luo, J., Sun, X., Cormack, B. P. and Boeke, J. D. (2018). Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature*, **560** (7718), 392–396.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R. *et al.* (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161** (5), 1012–1025.

Lupski, J. R. (1999). Charcot-marie-tooth polyneuropathy: duplication, gene dosage, and genetic heterogeneity. *Pediatric research*, **45** (2), 159–165.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends Genet.*, **19** (6), 330–338.

Marques-Bonet, T., Girirajan, S. and Eichler, E. E. (2009a). The origins and impact of primate segmental duplications. *Trends in Genetics*, **25** (10), 443–454.

—, Ryder, O. A. and Eichler, E. E. (2009b). Sequencing primate genomes: what have we learned? *Annu Rev Genomics Hum Genet*, **10**, 355–386.

Martin, C. L., Wong, A., Gross, A., Chung, J., Fantes, J. A. and Ledbetter, D. H. (2002). The evolutionary origin of human subtelomeric homologies—or where the ends begin. *The American Journal of Human Genetics*, **70** (4), 972–984.

Martin, J. S., Xu, Z., Reiner, A. P., Mohlke, K. L., Sullivan, P., Ren, B., Hu, M. and Li, Y. (2017). Hugin: Hi-c unifying genomic interrogator. *Bioinformatics*, **33** (23), 3793–3795.

McMahon, R., Sibbritt, T., Aryamanesh, N., Masamsetti, V. P. and Tam, P. P. (2021). Loss of foxd4 impacts neurulation and cranial neural crest specification during early head development. *Frontiers in cell and developmental biology*, **9**.

Mehrjouy, M. M., Fonseca, A. C. S., Ehmke, N., Paskulin, G., Novelli, A., Benedicenti, F., Mencarelli, M. A., Renieri, A., Busa, T., Missirian, C. *et al.* (2018). Regulatory variants of foxg1 in the context of its topological domain organisation. *European Journal of Human Genetics*, **26** (2), 186–196.

Merlo, G. R., Paleari, L., Mantero, S., Genova, F., Beverdam, A., Palmisano, G. L., Barbieri, O. and Levi, G. (2002a). Mouse model of split hand/foot malformation type I. *Genesis*, **33** (2), 97–101.

—, —, —, Zerega, B., Adamska, M., Rinkwitz, S., Bober, E. and Levi, G. (2002b). The dlx5 homeobox gene is essential for vestibular morphogenesis in the mouse embryo through a BMP4-mediated pathway. *Dev. Biol.*, **248** (1), 157–169.

Mertens, F., Johansson, B., Fioretos, T. and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, **15** (6), 371–381.

Meunier, J. and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.*, **21** (6), 984–990.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C. *et al.* (2012). A high-coverage genome sequence from an archaic denisovan individual. *Science*, **338** (6104), 222–226.

Miga, K. H. (2017). Chromosome-specific centromere sequences provide an estimate of the ancestral chromosome 2 fusion event in hominin genomes. *Journal of Heredity*, **108** (1), 45–52.

Mitelman, F., Johansson, B. and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, **7** (4), 233–245.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A. *et al.* (2021). Sustainable data analysis with snakemake. *F1000Research*, **10**.

Momozawa, Y. and Mizukami, K. (2020). Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, **66** (1), 11–23.

Muncke, N., Wogatzky, B., Breuning, M., Sistermans, E., Endris, V., Ross, M., Vetrie, D., Catsman-Berrevoets, C. and Rappold, G. (2004). Position effect on plp1 may cause a subset of pelizaeus-merzbacher disease symptoms. *Journal of medical genetics*, **41** (12), e121–e121.

Nazaryan-Petersen, L., Eisfeldt, J., Pettersson, M., Lundin, J., Nilsson, D., Wincent, J., Lieden, A., Lovmar, L., Ottosson, J., Gacic, J. *et al.* (2018). Replicative and non-replicative mechanisms in the formation of clustered cnvs are indicated by whole genome characterization. *PLoS genetics*, **14** (11), e1007780.

Ng, J., Sams, E., Baldridge, D., Kremitzki, M., Wegner, D. J., Lindsay, T., Fulton, R., Cole, F. S. and Turner, T. N. (2020). Precise breakpoint detection in a patient with 9p– syndrome. *Molecular Case Studies*, **6** (3), a005348.

Ning, Y., Rosenberg, M., Ledbetter, D. H. and Biesecker, L. G. (1996). Isolation of the human chromosome 22q telomere and its application to detection of cryptic chromosomal abnormalities. *Human genetics*, **97** (6), 765–769.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A. *et al.* (2022). The complete sequence of a human genome. *Science*, **376** (6588), 44–53.

—, Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M. and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, **30** (9), 1291–1305.

Oesper, L., Ritz, A., Aerni, S. J., Drebin, R. and Raphael, B. J. (2012). Reconstructing cancer genomes from paired-end sequencing data. In *BMC bioinformatics*, BioMed Central, vol. 13, pp. 1–13.

Ohno, S., Wolf, U. and Atkin, N. B. (2009). Evolution from fish to mammals by gene duplication. *Hereditas*, **59** (1), 169–187.

Ono, Y., Asai, K. and Hamada, M. (2021). Pbsim2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*, **37** (5), 589–595.

Pevzner, P. A. (1995). Dna physical mapping and alternating eulerian cycles in colored graphs. *Algorithmica*, **13** (1-2), 77–105.

Pienkowski, V. M., Kucharczyk, M., Młynek, M., Szczałuba, K., Rydzanicz, M., Poszewiecka, B., Skórka, A., Sykulski, M., Biernacka, A., Koppolu, A. A. *et al.* (2019). Mapping of breakpoints in balanced chromosomal translocations by shallow whole-genome sequencing points to efna5, bahd1 and ppp2r5e as novel candidates for genes causing human mendelian disorders. *Journal of Medical Genetics*, **56** (2), 104–112.

—, —, Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., Stawiński, P., Pollak, A., Kosińska, J., Wojciechowska, K., Lejman, M., Cieślikowska, A., Wicher, D., Stembalska, A., Matuszewska, K., Materna-Kiryluk, A., Gambin, A., Chrzanowska, K., Krajewska-Walasek, M. and Płoski, R. (2020). Breakpoint mapping of symptomatic balanced translocations links the EPHA6, KLF13 and UBR3 genes to novel disease phenotype. *Journal of Clinical Medicine*, **9** (5), 1245.

Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J. and Wishart, D. (2019). Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites*, **9** (4).

Poliak, S., Salomon, D., Elhanany, H., Sabanay, H., Kiernan, B., Pevny, L., Stewart, C. L., Xu, X., Chiu, S.-Y., Shrager, P., Furley, A. J. W. and Peles, E. (2003). Juxtaparanodal clustering of shaker-like k+ channels in myelinated axons depends on caspr2 and TAG-1. *J. Cell Biol.*, **162** (6), 1149–1160.

Poszewiecka, B., Gogolewski, K., Stankiewicz, P. and Gambin, A. (2022a). Revised time estimation of the ancestral human chromosome 2 fusion. *BMC Genomics*, **23** (S6).

—, Pienkowski, V. M., Nowosad, K., Robin, J. D., Gogolewski, K. and Gambin, A. (2022b). Tadeus2: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3d chromatin structure. *Nucleic Acids Research.*

—, Stankiewicz, P., Gambin, T. and Gambin, A. (2018). TADeus-a tool for clinical interpretation of structural variants modifying chromatin organization. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE.

Raaum, R. L., Sterner, K. N., Noviello, C. M., Stewart, C. B. and Disotell, T. R. (2005). Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *J. Hum. Evol.*, **48** (3), 237–257.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T. (2018). High-resolution tads reveal dna sequences underlying genome organization in flies. *Nature communications*, **9** (1), 189.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L. *et al.* (2015). Clingen—the clinical genome resource. *New England Journal of Medicine*, **372** (23), 2235–2242.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J. J., Kelso, J., Slatkin, M. and P??bo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468** (7327), 1053–1060.

Requena, F., Abdallah, H. H., García, A., Nitschké, P., Romana, S., Malan, V. and Rausell, A. (2021). CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients. *Nucleic Acids Research*, **49** (W1), W93–W103.

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S. and Martin, C. L. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genetics in Medicine*, **22** (2), 245–257.

Roberto, R., Misceo, D., D'Addabbo, P., Archidiacono, N. and Rocchi, M. (2008). Refinement of macaque synteny arrangement with respect to the official rhemac2 macaque sequence assembly. *Chromosome research*, **16** (7), 977–985.

Romiguier, J. and Roux, C. (2017). Analytical Biases Associated with GC-Content in Molecular Evolution. *Front Genet*, **8**, 16.

Ruan, J. and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods*, **17** (2), 155–158.

Satta, Y., Hickerson, M., Watanabe, H., O'hUigin, C. and Klein, J. (2004). Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. *J. Mol. Evol.*, **59** (4), 478–487.

Sawyer, S., Hartley, T., Dyment, D., Beaulieu, C., Schwartzentruber, J., Smith, A., Bedford, H., Bernard, G., Bernier, F., Brais, B. *et al.* (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clinical genetics*, **89** (3), 275–284.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C. and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483** (7388), 169–175.

Schuy, J., Grochowski, C. M., Carvalho, C. M. and Lindstrand, A. (2022). Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends in Genetics*.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S. *et al.* (2020). Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature biotechnology*, **38** (9), 1044–1053.

Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L. L., Zhou, F., Xiao, S., Liu, L., Zeng, X., Zheng, H., Yang, C., Zhao, Z., Zhao, G., Zhou, J. Q., Xue, X. and Qin, Z. (2018). Creating a functional single-chromosome yeast. *Nature*, **560** (7718), 331–335.

Sherman, J. H., Karpinski, B. A., Fralish, M. S., Cappuzzo, J. M., Dhindsa, D. S., Thal, A. G., Moody, S. A., LaMantia, A. S. and Maynard, T. M. (2017). Foxd4 is essential for establishing neural cell fate and for neuronal differentiation. *Genesis*, **55** (6), e23031.

Shioura, A., Tamura, A. and Uno, T. (1997). An optimal algorithm for scanning all spanning trees of undirected graphs. *SIAM Journal on Computing*, **26** (3), 678–692.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B. and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, **38** (11), 1348–1354.

Šošić, M. and Šikić, M. (2017). Edlib: a c/c++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, **33** (9), 1394–1395.

Stankiewicz, P. (2016). One pedigree we all may have come from - did adam and eve have the chromosome 2 fusion? *Molecular Cytogenetics*, **9** (1).

—, Shaw, C. J., Withers, M., Inoue, K. and Lupski, J. R. (2004). Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Research*, **14** (11), 2209–2220.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A. *et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, **144** (1), 27–40.

Stone, A. C., Battistuzzi, F. U., Kubatko, L. S., Perry, G. H., Trudeau, E., Lin, H. and Kumar, S. (2010). More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **365** (1556), 3277–3288.

Strathern, J. N., Shafer, B. K. and McGill, C. B. (1995). DNA synthesis errors associated with double-strand-break repair. *Genetics*, **140** (3), 965–972.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B. *et al.* (2012). The accessible chromatin landscape of the human genome. *Nature*, **489** (7414), 75.

Tomescu, A. I. and Medvedev, P. (2016). Safe and complete contig assembly via omnitigs. In *Research in Computational Molecular Biology: 20th Annual Conference, RECOMB 2016, Santa Monica, CA, USA, April 17-21, 2016, Proceedings 20*, Springer, pp. 152–163.

Tomita, E., Tanaka, A. and Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical computer science*, **363** (1), 28–42.

Turleau, C., De Grouchy, J. and Klein, M. (1972). Chromosomal phylogeny of man and the anthropomorphic primates. (pan troglodytes, gorilla gorilla, pongo pygmaeus). attempt at reconstitution of the karyotype of the common ancestor. *Ann. Genet.*, **15** (4), 225–240.

Ventura, M., Catacchio, C. R., Alkan, C., Marques-Bonet, T., Sajjadian, S., Graves, T. A., Hormozdiari, F., Navarro, A., Malig, M., Baker, C. *et al.* (2011). Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome research*, **21** (10), 1640–1649.

—, —, Sajjadian, S., Vives, L., Sudmant, P. H., Marques-Bonet, T., Graves, T. A., Wilson, R. K. and Eichler, E. E. (2012). The evolution of african great ape subtelomeric heterochromatin and the fusion of human chromosome 2. *Genome Research*, **22** (6), 1036–1049.

Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty, M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. and Eichler, E. E. (2019a). Long-read sequence and assembly of segmental duplications. *Nature methods*, **16** (1), 88–94.

—, Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M. and Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, **376** (6588).

—, Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A. M., Concepcion, G. T., Kronenberg, Z. N., Munson, K. M., Baker, C., Sanders, A. D., Spierings, D. C., Lansdorp, P. M., Surti, U., Hunkapiller, M. W. and Eichler, E. E. (2019b). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*, **84** (2), 125–140.

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A. and Batzer, M. A. (2005). SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.*, **354** (4), 994–1007.

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., Koenig, B. A., Li, D., Marschall, T., McMichael, J. F., Novak, A. M., Purushotham, D., Schneider, V. A., Schultz, B. I., Smith, M. W., Sofia, H. J., Weissman, T., Flicek, P., Li, H., Miga, K. H., Paten, B., Jarvis, E. D., Hall, I. M., Eichler, E. E. and and, D. H. (2022). The human pangenome project: a global resource to map genomic diversity. *Nature*, **604** (7906), 437–446.

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M. N. K., Li, Y., Hu, M., Hardison, R., Wang, T. and Yue, F. (2018). The 3d genome browser: a web-based browser for visualizing 3d genome organization and long-range chromatin interactions. *Genome Biology*, **19** (1).

—, Zhao, Y., Bollas, A., Wang, Y. and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology*, **39** (11), 1348–1365.

Wasa, K. (2016). Enumeration of enumeration algorithms. *arXiv preprint arXiv:1605.05102*.

Weckselblatt, B., Hermetz, K. E. and Rudd, M. K. (2015). Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome research*, **25** (7), 937–947.

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A. and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature Communications*, **10** (1).

Wells, R. A., Germino, G. G., Krishna, S., Buckle, V. J. and Reeders, S. T. (1990). Telomere-related sequences at interstitial sites in the human genome. *Genomics*, **8** (4), 699–704.

Wienberg, J., Jauch, A., Lüdecke, H. J., Senger, G., Horsthemke, B., Claussen, U., Cremer, T., Arnold, N. and Lengauer, C. (1994). The origin of human chromosome 2 analyzed by comparative chromosome mapping with a DNA microlibrary. *Chromosome Research*, **2** (5), 405–410.

—, —, Stanyon, R. and Cremer, T. (1990). Molecular cytotaxonomy of primates by chromosomal in situ suppression hybridization. *Genomics*, **8** (2), 347–350.

Wong, A., Vallender, E. J., Heretis, K., Ilkin, Y., Lahn, B. T., Martin, C. L. and Ledbetter, D. H. (2004). Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics*, **84** (2), 239–247.

Wong, A. C., Shkolny, D., Dorman, A., Willingham, D., Roe, B. A. and McDermid, H. E. (1999). Two novel human rab genes with near identical sequence each map to a telomere-associated region: the subtelomeric region of 22q13. 3 and the ancestral telomere band 2q13. *Genomics*, **59** (3), 326–334.

Yan, B., Neilson, K. M. and Moody, S. A. (2009). foxd5 plays a critical upstream role in regulating neural ectodermal fate and the onset of neural differentiation. *Developmental biology*, **329** (1), 80–95.

—, — and — (2010). Microarray identification of novel downstream targets of foxd4l1/d5, a critical component of the neural ectodermal transcriptional network. *Developmental Dynamics*, **239** (12), 3467–3480.

Yang, D., Jang, I., Choi, J., Kim, M.-S., Lee, A. J., Kim, H., Eom, J., Kim, D., Jung, I. and Lee, B. (2018). 3div: A 3d-genome interaction viewer and database. *Nucleic acids research*, **46** (D1), D52–D57.

Yunis, J. J. and Prakash, O. (1982). The origin of man: A chromosomal pictorial legacy. *Science*, **215** (4539), 1525–1530.

Zare, F., Dow, M., Monteleone, N., Hosny, A. and Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC bioinformatics*, **18** (1), 1–13.

Zepeda-Mendoza, C. J., Ibn-Salem, J., Kammin, T., Harris, D. J., Rita, D., Gripp, K. W., MacKenzie, J. J., Gropman, A., Graham, B., Shaheen, R. *et al.* (2017). Computational prediction of position effects of apparently balanced human chromosomal rearrangements. *The American Journal of Human Genetics*, **101** (2), 206–217.

Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics*, **24** (R1), R102–R110.

Zhang, J., Chiodini, R., Badr, A. and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of genetics and genomics*, **38** (3), 95–109.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R. M., Chang, C. C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S. T., Zaranek, A. W., Ball, M., Bobe, J., Estep, P., Church, G. M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G. X., Schnall-Levin, M., Ordonez, H. S., Mudivarti, P. A., Giorda, K., Sheng, Y., Rypdal, K. B. and Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, **3** (1).